# Bayesian Item Response Theory: Statistical Inference and Power Analysis 

Jason W. Bodnar<br>Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations
Part of the Educational Assessment, Evaluation, and Research Commons, Education Policy Commons, and the Statistics and Probability Commons

## Recommended Citation

Bodnar, Jason W., "Bayesian Item Response Theory: Statistical Inference and Power Analysis" (2011). Dissertations. 350.
https://scholarworks.wmich.edu/dissertations/350

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at
ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.

# BAYESIAN ITEM RESPONSE THEORY: STATISTICAL INFERENCE AND POWER ANALYSIS 

by<br>Jason W. Bodnar

A Dissertation
Submitted to the Faculty of The Graduate College in partial fulfillment of the requirements for the Degree of Doctor of Philosophy
Department of Educational Leadership, Research and Technology
Advisor: Brooks Applegate, Ph.D.

Western Michigan University
Kalamazoo, Michigan
December 2011

# BAYESIAN ITEM RESPONSE THEORY: STATISTICAL INFERENCE AND POWER ANALYSIS 

Jason W. Bodnar, Ph.D.

Western Michigan University, 2011

The regulatory pharmaceutical approval process is flawed in that industry clinical trials (ICTs) are always powered for efficacy and rarely powered for safety. The key safety parameter is the adverse event (AE). This practice may result in efficacious products with confounded safety. An ICT's ability to be powered for detecting AE trends may improve patient safety. Therefore, this dissertation's purpose was to determine if power analysis resulted in feasible sample sizes for substantiating AE hypotheses.

AEs were modeled with three Bayesian 2PL IRT models. The unidimensional latent trait, transfusion-related AE, was modeled as a patient predisposition for experiencing an AE. Parametric and nonparametric inference and power analysis approaches were derived for paired IRFs. Analysis was based on 1,000 bivariate binomial simulations of 9 AE types for $n=30$ and 250 patients. 2-PL, 2-PL EX, and 2-PL MEX adhered to the multiple chain assumption. Parameter estimators were stationary after 25,000 and 15,000 Gibbs samplers (GS), respectively, for 2-PL and 2-PL EX, and serial autocorrelation was removed. Simulation results revealed that the 2-PL EX demonstrated reasonable model fit based on linear trapezoid and spline approximations of the exact area under paired IRFs. Bootstrap, jackknife, and partial
batch approaches were used for parametric and nonparametric inference. Optimal results occurred for the nonparametric bootstrap approach on the spline approximation. Superiority was expectedly not achieved. Equivalence ( $\Delta=10 \%$ ) was not statistically substantiated for $n=30$, but was for $n=250$. Coverage was achieved for all inference. The superiority IRT approach required 933 patients for $95 \%$ confidence and $80 \%$ power different from existing methods which required a minimum of 174,451 patients. The equivalence IRT approach required 60 patients for $\Delta=10 \%$. Existing methods required a minimum of 95 patients.

Simulations resulted in comparable IRFs. Inference correctly characterized relationships between IRFs. IRT sample sizes were smaller than existing methods, and were expectedly larger for superiority. Powering a study to differentiate comparable groups typically requires enormous sample sizes. Equivalence was a viable solution to superiority. The next step is to have the IRT inference incorporated into the ICT safety investigation of all pharmaceutical products.

## All rights reserved

## INFORMATION TO ALL USERS

The quality of this reproduction is dependent on the quality of the copy submitted.
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

UMI 3496351
Copyright 2012 by ProQuest LLC.
All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.


ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346

Ann Arbor, MI 48106-1346

Copyright by Jason W. Bodnar 2011

## ACKNOWLEDGMENTS

This dissertation is dedicated to my lovely and supportive wife Peggy. Without this love and support, I would not have been able to reach the finish line.

I would also like to thank each of the Committee members for their valuable insights and writing recommendations.

Jason W. Bodnar

## TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... ii
LIST OF TABLES ..... viii
LIST OF FIGURES ..... xi
CHAPTER
I. INTRODUCTION. ..... 1
Statement of Problem ..... 1
Research Questions ..... 9
Contributions of the Study ..... 11
Study Objectives ..... 11
Statistical Inference ..... 12
Power Functions ..... 14
Risk Prediction Algorithms ..... 16
Limitations of the Study ..... 19
Study Definitions ..... 20
II. LITERATURE REVIEW ..... 26
Introduction to Item Response Theory ..... 26
Two-Parameter Logistic IRT Model ..... 29
IRT Model Assumptions ..... 31
Mathematical Statistics ..... 33

## Table of Contents-Continued

## CHAPTER

Bayesian Theory ..... 34
Bayesian Item Response Theory Models ..... 38
Gibbs Sampling ..... 42
Clinical Objectives ..... 45
Superiority. ..... 45
Hypothesis Testing in IRT Application ..... 46
Equivalence ..... 50
Equivalence Margin Estimation ..... 51
Power Analysis ..... 53
Item Response Theory ..... 55
Paired Binomial Endpoints ..... 59
Type I and II Errors for Superiority ..... 61
Type I and II Errors for Equivalence ..... 68
Pharmacokinetics ..... 71
AUC Estimation ..... 72
Risk Prediction Algorithms ..... 75
Application of IRT to Adverse Events ..... 77
III. METHODOLOGY ..... 79
Data Sources for Study ..... 79
Design of Study ..... 80

# Table of Contents-Continued 

## CHAPTER

Simulation of Bivariate Binomial Data ..... 80
Adverse Event Data Patterns ..... 83
Research Questions ..... 85
Research Question 1 ..... 85
Research Question 2 ..... 93
Research Question 3 ..... 102
IV. RESULTS ..... 104
Distributional Forms of Historical Data ..... 104
Quality of Monte Carlo Simulations ..... 106
Comparison of AUC Approximations ..... 111
Research Question 1 Results ..... 112
2-PL IRT Model ..... 113
2-PL EX IRT Model ..... 127
2-PL MEX IRT Model ..... 139
Research Question 2 Results ..... 155
Superiority Analysis ..... 160
Equivalence Margin Evaluation ..... 165
Equivalence Analysis ..... 167
Analyses on AE Characteristics ..... 173
Research Question 3 Results ..... 183

# Table of Contents-Continued 

CHAPTER
Existing Methodology Superiority Power Analysis ..... 183
IRT Superiority Power Analysis ..... 184
Existing Methodology Equivalence Power Analysis ..... 185
IRT Equivalence Power Analysis ..... 185
Summary of Results ..... 186
Research Question 1 ..... 186
Research Question 2 ..... 189
Research Question 3 ..... 191
V. DISCUSSION AND CONCLUSIONS ..... 193
Discussion ..... 193
DIF versus EIF ..... 195
Advantages of Study ..... 196
New IRT Knowledge Base. ..... 197
Directions for Future Research ..... 200
Data Pattern Anomalies ..... 200
Goodness-of-Fit ..... 201
Bayesian Sampling Algorithms ..... 201
Nonlinear Measurement Error ..... 202
Concluding Remarks ..... 203
REFERENCES ..... 205

## Table of Contents-Continued

## APPENDICES

A. Additional Convergence Plots for 2-PL IRT Model........................................... 222
B. Additional Convergence Plots for 2-PL EX IRT Model..................................... 235
C. Additional Convergence Plots for 2-PL MEX IRT Model ................................. 248
D. 2-PL IRT Model: Goodness-of-Fit..................................................................... 261
E. 2-PL IRT Model: AE Predisposition Estimates................................................... 270
F. 2-PL EX IRT Model: Goodness-of-Fit ................................................................ 276
G. 2-PL EX IRT Model: AE Predisposition Parameter Estimates .......................... 285
H. 2-PL MEX IRT Model: Goodness-of-Fit............................................................ 291
I. 2-PL MEX IRT Model: AE Predisposition Parameter Estimates ....................... 300
J. 2-PL EX IRT Model: Final Estimates of IRT Parameters ................................... 306
K. 2-PL EX IRT Model: Superiority and Equivalence Analyses ............................ 318
L. CDC Case Report Forms .................................................................................... 355
M. SAS Code.......................................................................................................... 358
N. Thrombocytopenia Phase III Clinical Trials ....................................................... 360

## LIST OF TABLES

1. $2 \times 2$ Contingency Table for Binomial Events. ..... 60
2. Predictive Power of Algorithms. ..... 77
3. Historical Incidence of Transfusion-Related Adverse Events in United States. ..... 81
4. Patterns of Simulated Adverse Event Characteristics ..... 84
5. Item Locations for Adverse Event Characteristics. ..... 84
6. Starting Values for Gibbs Samplers ..... 86
7. Relative Frequency of Transfusion-Related Adverse Events ..... 106
8. Metrics of Bivariate Binomial Simulation Algorithm ..... 108
9. 2-PL: Autocorrelation of Discrimination by AE Type. ..... 122
10. 2-PL: Autocorrelation of Difficulty by AE Type ..... 122
11. 2-PL: Autocorrelation of AE Predisposition ( $\boldsymbol{\theta}$ ) ..... 123
12. 2-PL: Discrimination by AE Type ..... 126
13. 2-PL: Difficulty by AE Type ..... 127
14. 2-PL EX: Autocorrelation of Discrimination by AE Type ..... 136
15. 2-PL EX: Autocorrelation of Difficulty by AE Type. ..... 136
16. 2-PL EX: Autocorrelation of AE Predisposition ( $\boldsymbol{\theta}$ ) ..... 137
17. 2-PL EX: Discrimination by AE Type ..... 140
18. 2-PL EX: Difficulty by AE Type ..... 140
19. 2-PL MEX: Autocorrelation of Discrimination by AE Type ..... 149

## List of Tables-Continued

20. 2-PL MEX: Autocorrelation of Difficulty by AE Type ..... 149
21. 2-PL MEX: Autocorrelation of AE Predisposition ( $\boldsymbol{\theta})$ ..... 150
22. 2-PL MEX: Discrimination by AE Type ..... 153
23. 2-PL MEX: Difficulty by AE Type. ..... 153
24. Summary of Approximate Area Under IRFs by Treatment $(n=30)$ ..... 158
25. Summary of Approximate Area Under IRFs by Treatment $(n=250)$ ..... 159
26. Superiority Analysis: Linear Trapezoid Approximation $(n=30)$ ..... 161
27. Superiority Analysis: Spline Approximation $(n=30)$ ..... 161
28. Superiority Analysis: Linear Trapezoid Approximation ( $n=250$ ) ..... 163
29. Superiority Analysis: Spline Approximation $(n=250)$ ..... 163
30. Evaluation of Various Equivalence Margins ..... 166
31. Equivalence Analysis: Linear Trapezoid Approximation ( $n=30$ ) ..... 168
32. Equivalence Analysis: Spline Approximation ( $n=30$ ) ..... 169
33. Equivalence Analysis: Linear Trapezoid Approximation ( $n=250$ ) ..... 171
34. Equivalence Analysis: Spline Approximation ( $n=250$ ) ..... 172
35. Superiority and Equivalence Analysis: Data Pattern 1 AE Type 1 ..... 175
36. Superiority and Equivalence Analysis: Data Pattern 2 AE Location 3 ..... 178
37. Superiority and Equivalence Analysis: Data Pattern 2 AE Location 6 ..... 178
38. Superiority and Equivalence Analysis: Data Pattern 3 AE Location 1 ..... 180
39. Superiority and Equivalence Analysis: Data Pattern 3 AE Location 20 ..... 181
40. Superiority and Equivalence Analysis: Data Pattern 4 All AE Locations ..... 182

## List of Tables-Continued

41. Superiority Sample Size Requirements for Existing Methods............................... 184

## LIST OF FIGURES

1. Medical Diagnosis Structure of Strep Throat ..... 3
2. Example Receiver Operator Characteristic Curve ..... 18
3. Example Two-Parameter Logistic IRT Curves ..... 30
4. Example Pharmacokinetic Curve ..... 71
5. Example Trace Plot for 1000 Gibbs Sampler Iterations ..... 87
6. Incidence of Transfusion-Related Adverse Events ..... 107
7. Relative Frequency of Transfusion-Related Adverse Events ..... 107
8. 2-PL: Simulation Plots for the Discrimination Parameter ..... 110
9. 2-PL: Simulation Plots for the Difficulty Parameter ..... 110
10. 2-PL: Simulation Plots for the AE Predisposition ( $\boldsymbol{\theta}$ ) Parameter ..... 111
11. 2-PL: Trace and G-R Plot for Mean Discrimination $(n=30)$ ..... 115
12. 2-PL: Trace and G-R Plot for Mean Discrimination $(n=250)$ ..... 116
13. 2-PL: Trace and G-R Plot for Mean Difficulty $(n=30)$ ..... 116
14. 2-PL: Trace and G-R Plot for Mean Difficulty $(n=250)$ ..... 117
15. 2-PL: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30)$ ..... 117
16. 2-PL: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=250)$ ..... 118
17. 2-PL: Trace and G-R Plot for $S D$ Discrimination $(n=30)$ ..... 118
18. 2-PL: Trace and G-R Plot for $S D$ Discrimination $(n=250)$ ..... 119
19. 2-PL: Trace and G-R Plot for $S D$ Difficulty $(n=30)$ ..... 119

## List of Figures-Continued

20. 2-PL: Trace and G-R Plot for $S D$ Difficulty $(n=250)$ ..... 120
21. 2-PL: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=30)$ ..... 120
22. 2-PL: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=250)$. ..... 121
23. 2-PL: Bayesian PPI and Residual Plot for AE $1(n=30)$ ..... 125
24. 2-PL: Bayesian PPI and Residual Plot for AE $1(n=250)$ ..... 125
25. 2-PL EX: Trace and G-R Plot for Mean Discrimination $(n=30)$ ..... 129
26. 2-PL EX: Trace and G-R Plot for Mean Discrimination $(n=250)$ ..... 129
27. 2-PL EX: Trace and G-R Plot for Mean Difficulty $(n=30)$ ..... 130
28. 2-PL EX: Trace and G-R Plot for Mean Difficulty $(n=250)$ ..... 130
29. 2-PL EX: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30)$ ..... 131
30. 2-PL EX: Trace and G-R Plot for Mean AE Predisposition ( $\boldsymbol{\theta}$ ) ( $n=250$ ) ..... 131
31. 2-PL EX: Trace and G-R Plot for $S D$ Discrimination $(n=30)$ ..... 133
32. 2-PL EX: Trace and G-R Plot for $S D$ Discrimination $(n=250)$ ..... 133
33. 2-PL EX: Trace and G-R Plot for $S D$ Difficulty $(n=30)$ ..... 134
34. 2-PL EX: Trace and G-R Plot for $S D$ Difficulty ( $n=250$ ) ..... 134
35. 2-PL EX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=30)$ ..... 135
36. 2-PL EX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=250)$ ..... 135
37. 2-PL EX: Bayesian PPI and Residual Plot for AE $1(n=30)$ ..... 138
38. 2-PL EX: Bayesian PPI and Residual Plot for AE $1(n=250)$ ..... 139
39. 2-PL MEX: Trace and G-R Plot for Mean Discrimination $(n=30)$ ..... 142
40. 2-PL MEX: Trace and G-R Plot for Mean Discrimination ( $n=250$ ) ..... 142

## List of Figures-Continued

41. 2-PL MEX: Trace and G-R Plot for Mean Difficulty $(n=30)$ ..... 143
42. 2-PL MEX: Trace and G-R Plot for Mean Difficulty $(n=250)$ ..... 143
43. 2-PL MEX: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30)$ ..... 144
44. 2-PL MEX: Trace and G-R Plot for Mean AE Predisposition ( $\boldsymbol{\theta})(n=250)$ ..... 144
45. 2-PL MEX: Trace and G-R Plot for $S D$ Discrimination $(n=30)$ ..... 146
46. 2-PL MEX: Trace and G-R Plot for $S D$ Discrimination $(n=250)$ ..... 146
47. 2-PL MEX: Trace and G-R Plot for $S D$ Difficulty $(n=30)$ ..... 147
48. 2-PL MEX: Trace and G-R Plot for $S D$ Difficulty $(n=250)$ ..... 147
49. 2-PL MEX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=30)$ ..... 148
50. 2-PL MEX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=250)$ ..... 148
51. 2-PL MEX: Bayesian PPI and Residual Plot for AE $1(n=30)$. ..... 152
52. 2-PL MEX: Bayesian PPI and Residual Plot for AE $1(n=250)$ ..... 152
53. IRT Plots for Fixed Historical Rate Data $(n=30)$ ..... 157
54. IRT Plots for Fixed Historical Rate Data $(n=250)$ ..... 164
55. PPI Plot for Data Pattern 1 AE Location 1 ..... 174
56. PPI Plot for Data Pattern 2 AE Location 3 ..... 177
57. PPI Plot for Data Pattern 2 AE Location 6 ..... 177
58. PPI Plot for Data Pattern 3 AE Location 1 ..... 179
59. PPI Plot for Data Pattern 3 AE Location 20 ..... 179
60. PPI Plot for Data Pattern 4 Across All AE Locations ..... 182

## CHAPTER I

## INTRODUCTION

## Statement of Problem

The International Conference on Harmonization (ICH) provides guidance for designing and conducting all aspects of industry clinical trials (ICTs) intended to support the regulatory approval of pharmaceutical products. Several of these guidance documents pertain to the evaluation of product safety (ICH E1, 1994; ICH E2a, 1995; ICH E3, 1995; ICH E6, 1996; ICH E8, 1997; ICH E9, 1998; ICH E11, 2000). If a regulatory agency agrees with a pharmaceutical company that their product's benefits outweigh its risks (i.e., safety is comparable to existing treatments and efficacy is statistically substantiated; ICH E9, 1998), the agency is thus indicating that the product is safe for market consumption.

A "safe" medical product is not one in which there is $0 \%$ risk to the consumer. No matter how safe a product is thought to be, at some point, a consumer will adversely respond to its consumption. The interaction of patient disease state and human genetics (Veenstra et al., 2007), concomitant medications (Jagsi et al., 2005), environmental factors (patient, hospital, etc.), and "outdated equipment" (Cowen \& Moorhead, 2006, p. 273) further complicate the occurrence, prediction, and prevention of such responses.

The safety of a medical product is gauged by its propensity for causing adverse events (AE). An AE is a diagnosis of a medical condition "temporally associated with the
use of a medicinal (investigational) product" (ICH E6, 1996, p. 7). These signs and symptoms may be based on objective measures or subjective observations, both of which may have binomial, ordinal, or continuous levels of measurement. Objective diagnostic criteria may comprise, but are not limited to, patient vital signs, laboratory assays, and radiology findings. Subjective diagnostic criteria such as scales (e.g., Visual Analog Scale for pain), self-reported experiences by the patient (e.g., nausea with no vomiting), or observations from a physician (e.g., facial flushing with no patient discomfort) may also lead to the reporting of an AE (ICH E3, 1995). When diagnosing a medical condition, physicians are typically concerned with a totality of clinical signs and symptoms.

Strep throat is an example of an AE that is diagnosed with both objectively and subjectively measured clinical signs and symptoms (Figure 1, p. 3). Per the Mayo Clinic (2011), this AE may be diagnosed with 5 patient-reported and 4 physician-observed signs and symptoms and 1 objective measure. The patient-reported signs and symptoms are throat pain, swallowing difficulty, headache, stomachache, and fatigue. The physicianobserved signs and symptoms are tonsils that are red and swollen, soft or hard palate having small red spots, swollen lymph nodes in the neck, and a rash. The last clinical sign and symptom is a fever, which is objectively measured with patient body temperature. All of these signs and symptoms are required for the diagnosis of strep throat per the Mayo Clinic criteria. The absence of any of these signs and symptoms is likely diagnostic of a different type of AE.

Risk Prediction Algorithms (RPAs) (Geraci, Rosen, Ash, McNiff, \& Moskowitz, 1993) use unique sets of clinical signs and symptoms to predict the occurrence of AEs at the patient level. After an RPA is developed for a specific clinical indication, medical


Figure 1. Medical Diagnosis Structure of Strep Throat
researchers may then fine-tune these algorithms by evaluating the reliability of the observable measures in terms of AE prediction. For ICTs, these algorithms may then be extended to AE characteristics (e.g., severity, seriousness, relation to investigational investigational treatment, etc.). For these trials, regulatory agencies require physicians to classify these characteristics for each type of AE that is reported. These agencies may then use these data in an attempt to further understand the safety implications associated with the use of the product. These assessments are conducted to identify any AEs that are causally associated with a particular AE, or whether AEs are a result of other medications the patient is taking, environmental factors, and/or patient genetics. The causal association may be classified by a physician with the categories ruled out, doubtful, possible, probable, or definite.

A factor that may further complicate this assessment is that some patients may have a greater predisposition to experiencing a particular type of AE than others (Empey, 2010). In essence, the treatment effect may interact with patient factors extraneous to the clinical trial. Furthermore, the type of AEs reported for a given population and treatment regimen is typically restricted to the context in which they occur (Battles, Kaplan, Van der Schaaf, \& Shea, 1998). Aspirin, for example, may be used to treat headaches experienced by healthy individuals or serve as a preventive measure for at-risk individuals predisposed to cardiac arrest. As a result, the types of AEs experienced by these two populations may differ. A person in the latter group, for instance, may be highly predisposed to developing an arrhythmia that could result in death (Drugs, 2011), whereas individuals in the former population may have a lower predisposition to this AE.

Given the complexities of investigating AEs in an ICT, an important question arises. How many patients are needed to develop a clinically sound understanding of medical product safety? ICH guidance recommends that 100 patients per year for a total of 1,500 patients over the product's development cycle should be exposed to the investigative treatment (ICH E1, 1994) for this purpose. Coincidentally, not all products approved by regulatory agencies are based on such sample sizes. As an example, the website http://clinicaltrials.gov indicates that 29 of 45 Phase III ICTs enrolled less than 100 patients for each group treated for thrombocytopenia or low platelet count (see Appendix N).

The reasons are relatively simple as to why Phase III trials may not employ the above sample size recommendations. First, the guidance does not statistically substantiate these sample sizes. As a result, the number of patients needed to understand the safety of
a medical product could vary greatly from study to study, product to product, or population to population. Sample sizes for these comparisons are commonly very large and often are not financially feasible. Second, Phase III trials are powered for the number of patients required to achieve efficacy or substantiating that the product works as medically intended (Chow, Shao, \& Wang, 2008).

For most ICTs, the primary objective is based on an efficacy variable.
Considerations regarding study design, hypothesis sets, sample size determination, and data analysis are based on this variable. Therefore, the power is properly computed for an analysis method used to evaluate efficacy hypotheses. This is not so for safety variables, such as the AE. This variable is frequently evaluated as a secondary objective in ICTs. Therefore, the power needed to substantiate AE hypotheses is not considered. In this case a statistical analysis method that is applied to the AE is considered as descriptive only. Mathematically, effects for efficacy are demonstrated by design, and any statistical inference demonstrated for AEs occurs by chance.

Powering a trial for efficacy and not AEs may result in products on the market that work as medically intended but are later found to have safety problems. The following statement by the Food and Drug Administration (1999) within the U.S. Department of Health and Human Services exemplifies this situation.

Once medical products are on the market, however, ensuring safety is principally the responsibility of healthcare providers and patients, who make risk decisions on an individual, rather than a population, basis. They are expected to use the labeling information to select and use products wisely, thereby minimizing AEs. (p. 4)

Studies powered on both efficacy and AE variables would provide a more thorough understanding of the efficaciousness and safety of approved products.

Substantiating that a product is both statistically efficacious and that the rate of AEs reported for a new treatment is statistically and clinically comparable to an existing treatment prior to market release would logically be beneficial to the general public.

This study compared sample size requirements for AEs generated between the current statistical methodology and the newly introduced method based on Item Response Theory (IRT; Lord, 1953). If the sample size requirements were found to be comparable between the IRT and existing methods, this result would reinforce the difficulty of powering an ICT on AE hypotheses. On the other hand, if the IRT-based sample sizes were comparable to efficacy requirements, the pharmaceutical industry may have access to a new statistical tool that could revolutionize the evaluation of product safety.

IRT differs from classical statistical methods in that it is a latent trait model approach (Hambleton \& van der Linden, 1982). This model uses mixed-effects regression (Hedeker, Mermelstein, \& Flay, 2006) to estimate unobservable construct or latent trait parameters (Hambleton \& Cook, 1997; Junker, 1997). IRT may be used to understand the "relationship between observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test" (Hambleton \& Cook, 1977, p. 75).

In this study, the latent trait was referred to as the Transfusion-Related Adverse Event. This trait was an unobservable construct that represented the medical diagnosis of a unique set of one or more observable clinical signs and symptoms. Each type of AE or item was measured with the binomial response of either a 1 or 0 . In a traditional IRT application, these values may mean, respectively, that a student passed or failed a test
item. In this study, these values, respectively, indicated that a unique set of clinical signs and symptoms were or were not diagnostic of a particular type of AE.

This study attempted to expand the use of this diagnostic information for two types of patients. The first type of patient is those who are enrolled in an ICT and were anticipated to experience AEs, but this experience has not yet occurred. The second group consisted of those patients who signed an informed consent form so that they could participate in an ICT but their enrollment has not yet begun. A common factor associated with these two groups of patients is that they have not experienced any clinical signs and symptoms that would be diagnostic of a particular type of AE.

Instead of evaluating the diagnosis of AEs, this study developed IRT methodology for modeling the predisposition of patients being diagnosed for a particular type of AE or AE Predisposition ( $\boldsymbol{\theta})$. Disease states, human genetics, concomitant medications, environmental factors, and old equipment can alter this predisposition for a given investigational treatment. Furthermore, because this treatment alone is unlikely to cause the same AE for all patients enrolled in an ICT, the occurrence of this AE can be better understood in relation to extraneous patient factors. Various combinations of these factors may result in the differential occurrences of clinical signs and symptoms that are diagnostic of a particular type of AE. That is, certain patients may be more or less predisposed for experiencing AEs.

Two-parameter logistic IRT (Embretson \& Reise, 2000) was used to model the ability parameter $(\boldsymbol{\theta})$ based on the latent trait Transfusion-Related Adverse Event. For the remainder of this dissertation, this parameter is referred to as "AE Predisposition $(\boldsymbol{\theta})$." The parameters used to characterize AE Predisposition ( $\boldsymbol{\theta})$ for $j$ types of AEs were
referred to as item discrimination and difficulty (Swaminathan, Hambleton, Sireci, Xing, \& Rizavi, 2003). The discrimination parameter $a_{j}$ (for $\mathrm{j}^{\text {th }} \mathrm{AE}$ ) was used to estimate the magnitude of the relationship between $j$ items and the latent trait (Reeve \& Fayers, 2005). This parameter characterized the differential contribution of each item on the latent trait. Moreover, these AE items were differentiated by type (e.g., infection) and item-specific characteristics such as severity, relationship to treatment, and seriousness. The parameter $b_{j}$ represented the difficulty (Amarnani, 2009) or observed occurrence of $j$ types of AEs. Furthermore, the estimated parameters $\boldsymbol{\theta}_{i}, a_{j}$, and $b_{j}$, could be used to determine individual patient probabilities of experiencing an AE. This probability may be calculated with Equation 1, the IRT version of the logistic cumulative density function (CDF).

$$
\begin{equation*}
P\left(\theta_{i} ; a_{j}, b_{j}\right)=e^{-a_{j}\left(\theta_{i}-b_{j}\right)} /\left(1+e^{-a_{j}\left(\theta_{i}-b_{j}\right)}\right) \tag{Eq.1}
\end{equation*}
$$

These latent trait parameters were investigated for two study objectives commonly evaluated in ICTs: superiority and equivalence (ICH E9, 1998). Superiority is substantiated if one type of treatment is statistically better than or different from another type of treatment on a response variable (Wiens, 2006). IRT applications have utilized this objective but it is typically referred to as Differential Item Functioning (DIF; Bock, 1997). Uses of superiority for IRT application will be presented in Chapter II.

Equivalence is achieved if one treatment group is statistically comparable to another treatment group on a response variable (Lui, 2005). This study attempted to incorporate equivalence into IRT application.

Statistical inference (i.e., hypothesis testing and confidence intervals) was derived by the author for superiority and equivalence study objectives to be evaluated with IRT
statistical methodology. The estimator for this inference was the approximated area under Item Response Functions (IRFs) for the parameter AE Predisposition ( $\boldsymbol{\theta}$ ) for each type of AE. The mathematical definitions of the effects and their components will be derived by the author in Chapter III. A paired data structure, common to transfusion medicine, was employed for these IRFs. In a traditional IRT setting, this data structure may mimic a student taking two exams of the same design. This study assumed that each patient received both Treatment A and Treatment B, which represented a new treatment and an existing treatment regimen, respectively.

Functions used for power analysis will next be derived by the author in Chapter III. These functions were used to determine the likelihood or probability of demonstrating a particular objective for a given sample size. Specifically, sample size estimation was used to determine the number of patients required to demonstrate superiority or equivalence between treatment groups on the AE Predisposition $(\boldsymbol{\theta})$ parameter.

## Research Questions

This study utilized simulated data that coincide with the historical incidence of AEs occurring from the transfusion of blood products (U.S. Department of Health and Human Services, 2007b). These products are but are not limited to whole blood and its nested components red blood cells, platelets, and plasma. The research questions explored with these data were:

1. Can one or more two-parameter Bayesian logistic IRT models be used to model AEs? This feasibility was evaluated with statistics used to assess Gibbs sampling algorithms and goodness-of-fit (GoF). IRT was considered a viable solution for
modeling AEs under the study constraints if these algorithms demonstrated convergence and GoF was established.
2. For superiority and equivalence study objectives, which combination of statistics resulted in the best statistical inference as defined by
a. minimum standard error of the effect?
b. smallest bias in Area-Under-the-Curve (AUC) approximations?
c. confidence interval on the effect that achieves the highest coverage?

Several combinations of statistical inference and its components on paired IRFs for both superiority and equivalence study objectives were investigated in this study. The combination of statistics that resulted in the best statistical inference was recommended for utilization.
3. How do IRT sample size requirements for superiority and equivalence study objectives compare to the existing classical methods for paired binomial variables?

This question pertained to the concept of utility. Financially feasible sample sizes (i.e., those comparable to efficacy requirements) will be necessary in order for pharmaceutical companies to adopt and incorporate the presented IRT methodology into their ICTs. Otherwise, IRT may only be used as a secondary analytic method.

## Contributions of the Study

## Study Objectives

This study resulted in four key contributions that attempt to advance the inferential investigation of rare binary events with IRT methodology. The first contribution was the incorporation of superiority and equivalence study objectives into IRT. Superiority enables the determination of whether or not a variable for one treatment group is better than or different from another group. Equivalence enables the determination of whether or not a variable for one treatment group is statistically the same as another group. An advantage of these objectives is that they are not restricted to any particular type of IRT model.

Paired IRFs for treatments A and B were investigated for these objectives. The paired or $k$-sample matched design was employed because it is commonly implemented in transfusion medicine ICTs (Fergusson, Hebert, \& Shapiro, 2002; Hshieh \& Ng, 2007). Objectives were analyzed for the AE Predisposition ( $\boldsymbol{\theta})$ parameter. If superiority was statistically substantiated this would mean that the predisposition of a patient experiencing an AE (overall or individual) was better than (i.e., lower) in Treatment A than Treatment B. If equivalence was statistically substantiated this would mean that the predisposition of patients experiencing an AE (overall or individual) was the same between treatment groups.

## Statistical Inference

The next contribution consisted of the derivation of complete statistical inference by the author for the $k$-sample matched design for the presented IRT statistical methodology. Formal hypothesis sets were developed for both superiority and equivalence study objectives. A parametric and nonparametric solution was then presented for the effect and its components (i.e., standard error, test statistic, and confidence interval) for these study objectives. The estimator of the effect was based on the approximate area under paired IRFs. The AUC statistic was utilized because it does not have units nor requires a known distribution.

Linear trapezoid and spline approximations to the exact area under paired IRFs were computed. The linear trapezoid approximation assumes that the distance between any two sets of points is connected with a linear function (Chiou, 1978). The spline approximation assumes that the distance between any three sets of points is connected with a cubic function (Yeh \& Kwan, 1978). It was anticipated that the two approximations would be comparable for linear functions, and the spline approximation would be less biased for nonlinear functions (Yeh \& Kwan, 1978), such as the logistic which generally has a cubic polynomial form (Hogg, McKean, \& Craig, 2005).

Furthermore, the Partial (Bandos, Rockette, \& Gur, 2007) and Bailer (1988) approaches were utilized to estimate the area under paired IRFs across all types of AEs. Due to volume of output, if these aggregation approaches reached comparable estimates, results would be limited to the Partial approach. For each type of AE, the Partial Batch (Navarro-Fontestad, González-Álvarez, Fernández-Terul, Bermejo, \& Casabó, 2005),

Bootstrap (Dunning, 2007), and Jackknife (Zou, Gastwirth, \& McNeil, 2003) approaches were used for estimation. A key advantage of the above statistical methodology is that it is universal to IRT models concerned with response functions (item and person) regardless of variable level of measurement.

The derived statistical inference may also impact how hypotheses are used in IRT application. As an example, DIF, which is currently only based on a superiority objective, may be used to determine if groups differ on a latent trait (Bolt, 2000; Caufmann \& MacIntosh, 2006; Glickman, Seal, \& Eisen, 2009; Swaminathan \& Rogers, 1990). The general hypothesis set for this objective can be portrayed as
$\mathrm{H}_{0}$ : No DIF
$\mathrm{H}_{1}$ : DIF
DIF can only provide evidence as to whether groups behave differently on a latent trait. Hays, Morales, and Reise (2000) further indicate that DIF is present when the probability of two or more groups is different for endorsing a latent trait item. Such significance would be characterized by the lower confidence limit on the probability being greater than zero (Bolt, 2000). If this confidence limit was negative, a claim of DIF would not be statistically warranted.

In this situation, an alternative study objective such as equivalence may be more appropriate. This objective, loosely speaking, can be thought of as the antithesis of superiority. That is, equivalence may be portrayed as
$\mathrm{H}_{0}$ : DIF
$\mathrm{H}_{1}$ : No DIF

For this hypothesis set, a significant $p$-value would indicate that groups are invariant or comparable on a latent trait. This study strived to make the first attempt at utilizing equivalence for demonstrating lack of DIF (Cook et al., 2007). The Equivalent Item Function (EIF) was developed as a counterpart to DIF.

## Power Functions

The third contribution arose from the development of parametric and nonparametric power functions for both superiority and equivalence study objectives. These functions provided the a priori probability of demonstrating an alternative hypothesis for an objective for a given sample size. In this study, these functions were used to determine sample size requirements for an a priori probability of at least 0.80 with $100(1-\alpha) \%$ confidence of statistically substantiating objectives across all AE types. Superiority was concerned with demonstrating that treatment groups were statistically different in AE Predisposition $(\boldsymbol{\theta})$. Equivalence was concerned with demonstrating that treatment groups were statistically comparable in this parameter. A key advantage of the power functions derived by the author is that they are universal to IRT models concerned with response functions (item or person) regardless of variable level of measurement.

Two additional advantages to these power functions are noteworthy. Although this study focused on patient-level sample size requirements for comparing AE Predisposition $(\boldsymbol{\theta})$, these functions are equally applicable to items in terms of random treatment allocation. The patient was the unit of analysis used in this study because the number and type of AEs cannot be controlled in an ICT. Outside of this setting, it may be possible for items to be selectively modified (Bolt, 2000) or removed (Glas \& Meijer, 2003) from a
test instrument or survey, for example. Second, these functions and their statistical inference are readily adaptable to the $k$-sample independent study design (MaydeuOlivares, 2005). This adaptation can be performed by setting all covariances equal to zero and changing the degrees of freedom from $n-1$ to $2(n-1)$, where $n$ represents the number of units of analysis (e.g., patients). In certain settings such as psychological and educational testing, independent "treatment" groups may be necessary to avoid difficulties such as "learning effects" (Haynie, 2007).

Swaminathan, Hambleton, Sireci, Xing, and Rizavi (2003) claim that sample size estimation for IRT "has been well studied" (p. 2). This conclusion may be a result of many articles presenting sample size recommendations based on informal rules for IRT application (Bock, 1997; Cepicka, 2003; Embretson \& Reise, 2000; Linacre, 1994; Reeve \& Fayers, 2005; Teresi, Kleinman, \& Ocepek-Welikson, 2000). However, Holman, Math, Glas, and de Haan (2003) claim that "minimal sample size and power calculations in relation to questionnaires analyzed with IRT have received very little attention" (p. 391). These authors further add that "no guidance on sample size calculations for RCTs in the context of IRT has been published" (p. 391).

The Western Michigan University library system, WorldCat, Web of Science, ERIC, and $\mathrm{ABI} /$ Inform document databases were searched for all combinations of the terms item response theory, power function, and power analysis. The results returned a total of 9 articles from all of these searches, which will be detailed in Chapter II. Five of the articles specified that power analysis was performed, but possessed insufficient detail for sample size reproduction. The remaining articles presented formal power functions or steps used to compute power.

## Risk Prediction Algorithms

The last contribution concerned an additional use for IRT methodology in ICTs. AEs attributable to a product (Department of Veterans Affairs, 2008) that are serious (Rochester, 2003), unexpected, or uncommon (Rochester, 2009) are typically of most concern to "decision-makers and stakeholders" (Rochester, 2003). Regression models (Black, Markides, \& Ray, 2003; D’Agostino, Grundy, Sullivan, \& Wilson, 2001; DuMouchel, 2010; Lin, Hosmane, Olson, \& Padley, 2001; Nishtala et al., 2009) are commonly used to understand how various combinations of demographics, disease states and biomarkers, concomitant medications, and treatment regimens relate to the occurrence of AEs (Haq, Jackson, Yeo, \& Ramsay, 1995).

Two problems complicate using interaction effects to understand clinically important relationships between patient-level AEs and risk factors. First, these trials are typically powered only for main effects on efficacy. As a result, clinically important interaction effects are likely to be confounded with sample size. Interaction effect $p$ values are typically too large to be clinically meaningful for aggregate- or patient-level conclusions. A nonsignificant $p$-value may not necessarily mean that a certain combination of factors lacks predictive capability of a particular type of AE. Here, the problem could be a result of insufficient statistical power. Second, the relationship between the variables comprising the interaction effect is assumed to be linear. This will be problematic if a level of one main effect coincides with multiple levels of a second main effect.

The Risk Prediction Algorithm (RPA; Geraci et al., 1993) may be an alternative to the traditional interaction effect. As applied to this study, clinically relevant risk factor scores could be potentially used to develop gradations of AE Predisposition ( $\boldsymbol{\theta}$ ) for a specific medical condition. As the gradation increases in magnitude the susceptibility (Rochester, 2009) or predisposition of a patient experiencing an AE may increase. Because the use of RPAs for AEs is a recent development in ICTs, the form of this relationship has not been established.

An AE would be anticipated to occur for a particular type of patient once a certain gradation was reached. Physicians could then use this type of information to potentially reduce the predisposition to or severity of (Oberg, 1999) an AE by proactively (Chhibber \& King, 2010; Hanson \& France, 2004; Krediet, van Dijk, Linzer, van Lieshout, \& Weiling, 2002) treating a patient (i.e., providing alternative or modified treatment regimens). RPAs have been developed for a number of medical indications with risk factors specific to these conditions, excluding transfusion medicine, and these will be detailed in Chapter II.

In order for RPAs to have utility in medical settings, their accuracy or predictive utility (Bridgewater, Neve, Moat, Hooper, \& Jones, 1998) must be clinically sufficient. The Receiver Operator Characteristic curve (ROC) has commonly been utilized to quantify such utility. ROC curves are based on specificity and sensitivity, both of which range from 0 to 1 (Zhang \& Mueller, 2005). As applied to this study, specificity would be defined as the false positive rate of AE occurrence. Sensitivity would occur when an algorithm correctly predicts the presence of an AE. As sensitivity increases with specificity, the accuracy of the prediction algorithm increases. In theory, if the false
negative and positive rates are zero (i.e., $\mathrm{AUC}=1$ ), the algorithm would have $100 \%$ predictive ability.

An example of two ROC curves is provided in Figure 2. Curve 1 and Curve 2, respectively, demonstrate nonlinear and linear relationships between specificity and sensitivity. When sensitivity equals 0.9 , specificity approximately equals 0.58 and 0.93 , respectively, for Curve 1 and Curve 2. These results would indicate that when the algorithm correctly predicts the occurrence of $90 \%$ of AEs, the false positive rate is lower for Curve 1. As a result, Curve 1 has more predictive utility of the AEs under investigation. This type of methodology would be amendable to comparing the algorithms for two RPAs.


Figure 2. Example Receiver Operator Characteristic Curve

IRT may be a new statistical framework for developing and comparing RPAs. IRT would substitute patient-level (conditional) probabilities for gradations of risk scores. The likelihood of an AE occurring would increase with the resulting probability (de Ayala, 2009). As the width of the confidence interval of this probability decreases, the precision of the probability would increase. Furthermore, algorithms based only on probability may be easier to clinically interpret. A $90 \%$ chance of a patient with a certain set of risk factors experiencing a particular type of AE may be more informative than indicating that a risk score above a certain threshold is expected to be $90 \%$ accurate. Regardless, ROCs (Hanley \& McNeil, 1982) via $c$ statistics and net reclassification improvement methods (Steyerberg et al., 2010) will be required to re-evaluate and fine-tune the predictive ability of an IRT-based RPA.

## Limitations of the Study

Three limitations to this study are relevant. First, this study was limited to IRT models for latent traits measured by items with binomial responses. Other levels of measurement (e.g., ordinal shift tables for laboratory assays) are infrequently used in ICTs. Many IRT applications in academic "clinical research, psychology, educational sciences, ecology, and epidemiology" (Hardouin, 2007, p. 1) may use variables with ordinal and continuous levels of measurement. Although the statistical inference and power analysis methodology presented in this study is universal to IRT models concerned with response functions (item and person), it is unknown which study design will be optimal in each setting until appropriate investigation is performed.

Second, the predisposition for a patient experiencing an AE was modeled with the standard normal probability density function (PDF). This distribution resulted in the latent trait scores having range space $\theta \in[-3,+3]$ (Thompson, 2009). It is unknown if this distribution can be used to sufficiently model the ability parameter $(\boldsymbol{\theta})$ in all settings (Samejima, 1997) such as ICTs. This study assumed this optimality for AE Predisposition $(\boldsymbol{\theta})$. The primary reason for this decision surrounded the interpretability of this parameter. If a different distribution was assumed, the range space of this parameter may not be $[-3,+3]$. In this case, the interpretation of this parameter may change. For example, if the Beta PDF was used to model this parameter, its range space would change to $(0,1)$.

Last, de Ayala (2009) indicates that self-reported experiences are latent traits. During the course of an ICT, the patient or physician may be the source of multiple types of AEs. Factor analysis techniques would be required to determine if these AE types are distinct latent traits. If this analysis substantiated the presence of multiple latent traits, it would be anticipated that an IRT model such as the multidimensional two-parameter logistic (McKinley \& Way, 1992) would possess inferential advantages over the single dimension counterpart. Because the dimensionality of AEs is not understood, this study was restricted to a single latent trait, Transfusion-Related Adverse Event.

Study Definitions

Adverse Event (AE) - "An AE is any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and that does not necessarily have a causal relationship with this treatment. An AE can therefore be
any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product." (ICH E6, 1996, p. 7)

Clinically Significant AE - An AE that "causes harm or illness and/or requires testing, monitoring, or short-term or long-term treatment." (Department of Veterans Affairs, 2008, p. E-1)

Serious AE - "A serious adverse event (experience) or event is any untoward medical occurrence that at any dose

- results in death,
- is life-threatening, [Note: The term "life-threatening" in the definition of "serious" refers to an event in which the patient was at risk of death at the time of the event; it does not refer to an event which hypothetically might have caused death if it were more severe.]
- requires inpatient hospitalisation or prolongation of existing hospitalisation,
- results in persistent or significant disability/incapacity, or
- is a congenital anomaly/birth defect." (ICH E2A, 1995, p. 4)

Plasma - "Plasma is a fluid, composed of about $92 \%$ water, $7 \%$ vital proteins such as albumin, gamma globulin, anti-hemophilic factor, and other clotting factors, and 1\% mineral salts, sugars, fats, hormones and vitamins." (American Red Cross, 2010)

Red Blood Cells - "Red cells, or erythrocytes, carry oxygen from the lungs to your body's tissue and take carbon dioxide back to your lungs to be exhaled." (American Red Cross, 2010)

Platelets - "Platelets, or thrombocytes, are small, colorless cell fragments in the blood whose main function is to interact with clotting proteins to stop or prevent bleeding." (American Red Cross, 2010)

Random Donor Platelets - Platelets centrifuged from whole blood after donation of whole blood (Silberman, 1999).

Apheresis - "Automated blood collection in which a device continuously or intermittently removes a small volume of whole blood, separates the components, collects certain components, and returns to the donor the uncollected remainder." (U.S. Department of Health and Human Services, 2007a, p. 6)

Single Donor_Platelets - Platelets extracted from whole blood during blood component donation (Silberman, 1999).

Allergic Event (including anaphylaxis) - "The result of an interaction of an allergen with preformed antibodies. In some instances, infusion of antibodies from an atopic donor may also be involved." (Centers for Disease Control, 2010, p. 11)

Acute Hemolytic Transfusion Event - "Rapid destruction of red blood cells during, immediately after, or within 24 hours of a transfusion." (Centers for Disease Control, 2010, p. 12)

Delayed Hemolytic Transfusion Event - "The recipient develops antibodies to RBC antigen(s) between 24 hours and 28 days after a transfusion." (Centers for Disease Control, 2010, p. 13)

Delayed Serologic Transfusion Event - "Demonstration of new, clinically significant alloantibodies against red blood cells between 24 hours and 28 days after a transfusion despite an adequate, maintained hemoglobin response." (Centers for Disease Control, 2010, p. 14)

Febrile Nonhemolytic Transfusion Event - "Fever and/or chills without hemolysis occurring in the patient up to 4 hours during and after transfusion." (Centers for Disease Control, 2010, p. 16)

Hypotensive Transfusion Event - "A drop in systolic and/or diastolic blood pressure occurring during or within one hour of completing transfusion." (Centers for Disease Control, 2010, p. 15)

Infection (Bacterial - including sepsis / Viral / Organism) - "A bacteria, parasite, virus, or other potential pathogen transmitted in donated blood to transfusion recipient." (Centers for Disease Control, 2010, p. 22)

Transfusion Associated Circulatory Overload - "Infusion volume that cannot be effectively processed by the recipient either due to high infusion rate and/or volume or an underlying cardiac or pulmonary pathology." (Centers for Disease Control, 2010, p. 18)

Transfusion Associated Graft vs Host Disease - "The introduction of immunocompetent lymphocytes into susceptible hosts. The allogeneic lymphocytes engraft, proliferate and destroy host cells." (Centers for Disease Control, 2010, p. 20)

Probability Density Function (PDF) - A function, a mathematical representation of a concept, is classified as a PDF when the property $\int_{-\infty}^{\infty} f(x) d x=1$ holds. This
requirement means that if the total area under a function is 1 the function meets this property. If this property is met then inference can be developed for the random variable X using this function. (Hogg \& Craig, 1995)

Marginal PDF - The marginal PDF of X given the joint $\operatorname{PDF} \mathrm{f}(\mathrm{x}, \mathrm{y})$ is equal to

$$
f(x)=\int_{-\infty}^{\infty} f(x, y) d y(\text { Ross, 1997). }
$$

Conditional PDF - The conditional PDF of X given the joint $\operatorname{PDF} \mathrm{f}(\mathrm{x}, \mathrm{y})$ is

$$
\mathrm{f}(\mathrm{x} \mid \mathrm{y})=\frac{\mathrm{f}(\mathrm{x}, \mathrm{y})}{\mathrm{f}(\mathrm{y})} \text { (Ross, 1997). }
$$

Range Space - The range of values under which the data of a parameter behaves.
Normal PDF - The normal PDF is a distribution used for representing parameters that have nonlinear values between $-\infty$ to $+\infty$ (Hogg \& Craig, 1995). The functional form of this PDF is $f\left(x ; \mu, \sigma^{2}\right)=\frac{1}{\left(2 \sigma^{2} \pi\right)^{1 / 2}} \exp \left\{-\frac{(x-\mu)^{2}}{2 \sigma^{2}}\right\},-\infty<x, \mu<\infty, \sigma^{2}<$ $\infty$.

Beta PDF - A distribution used for representing parameters that have nonlinear values between 0 and 1 (Daly, 1992). The functional form of this PDF is

$$
f(x ; a, b)=\frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} x^{a-1}(1-x)^{b-1}, x \in(0,1), a, b>0 .
$$

Gamma PDF - A distribution used for representing parameters that have nonlinear values between 0 and $+\infty$ (Hogg \& Craig, 1995). The functional form of this PDF is

$$
f(x ; a, b)=\frac{1}{\Gamma(a) b^{a}} x^{a-1} e^{-x / b}, x \geq 0, a, b>0 .
$$

Cumulative Distribution Function (CDF) - The CDF equals the integral of a PDF with lower limit equal to the minimum range space and the upper limit equal to the random variable. The CDF has the general form $F(x)=\int_{-\infty}^{x} f(t) d t$, where $F(x)$ ranges from 0 to 1 (Ross, 1997).

## CHAPTER II

## LITERATURE REVIEW

This chapter reviews the literature associated with approaches to statistical inference and power analysis for comparing binomial variables between paired treatment groups for clinical objectives. This chapter begins with an introduction to Item Response Theory (IRT), and then discusses the forms of the IRT models to be explored in this study. A brief background is then provided on mathematical statistics to assist the understanding of Bayesian theory and its assumptions for IRT models. The chapter then discusses IRT application in terms of evaluating superiority and equivalence study objectives. Next, the literature on determining sample size requirements is provided for IRT models. Thereafter, the existing methods for computing Type I and II errors for paired binomial variables are presented. The chapter then provides background on and formulae for approximating the exact Area-Under-the-Curve (AUC) utilized for comparing paired IRFs. Next, RPAs, a potential new application for IRT and the means for constructing them, are discussed. This chapter ends with a discussion of existing IRT methodology for modeling adverse events (AEs).

Introduction to Item Response Theory

IRT is a nonlinear mixed-model regression approach (Hedeker, Mermelstein, \& Flay, 2006) for modeling data structures containing one or more latent traits. A latent trait
is an unobservable construct (Hambleton \& Cook, 1997; Junker, 1997) that is indirectly measured with ability or proficiency parameters (Reeve \& Fayers, 2005). Examples of this construct may include but are not limited to education (e.g., mathematics comprehension) (Fox \& Glas, 2001), psychology (e.g., anxiety), and medicine (e.g., headache) (de Ayala, 2009). The ability or proficiency parameter, also referred to as a manifest variable (de Ayala, 2009), may take the form of a standardized mathematics test to evaluate an educational latent trait, an anxiety scale to evaluate a psychological latent trait, or visual analog scales to evaluate a medical latent trait.

The level of measurement of these ability or proficiency variables then guides how the associated latent traits are statistically investigated. Logistic or Normal-Ogive models are commonly used for binomial responses (Embretson \& Reise, 2000). Rating Scale, Partial Credit, and Nominal Response models are appropriate for ordinal or polytomous data (van der Linden \& Hambleton, 1997). IRT can also be utilized to model variables that have a continuous level of measurement such as summed scale scores (Drachler, Marshall, \& de Carvalho, 2007) and numerical responses (Noel \& Dauvier, 2007).

For each of these types of IRT models, parameter estimation may be performed within one or more theoretical frameworks. Joint or Marginal Maximum Likelihood Estimation (Jannarone, Yu, \& Laughlin, 1990) utilizes an asymptotic approach to estimate parameter estimates. This parametric approach commonly assumes that a normal probability density function (PDF) underlies the distribution of the parameter estimates (Samejima, 1997). A statistical framework that does not assume a particular PDF for a data structure falls into the nonparametric category (de Leeuw \& Hox, 1998). The third
framework, Bayesian, is an extension of Maximum Likelihood Estimation (MLE), and overcomes a key limitation of the classical approaches. Unlike solutions from classical methods, Bayesian estimates can be computed for respondents who have null or perfect scores (Hardouin, 2007). A null and perfect score represents a respondent who failed and passed, respectively, all test items on an instrument such as an exam (Swaminathan \& Gifford, 1985). For this study, these scores represented a patient who experienced none or all of the AEs reported during an industry clinical trial (ICT).

IRT models, regardless of the theoretical framework, density assumptions, and the level of measurement for responses to items, generally aim to estimate four types of parameters. The parameter $\boldsymbol{Z}$ denotes the $n \times k$ matrix of unobservable latent trait values for each respondent $n$ and item $k$. This parameter's range space is dependent upon the assumed PDF. For example, this range may be -3 to +3 for a normal PDF. The ability parameter, $\boldsymbol{\theta}$, denotes the $n \times 1$ vector of ability values for $n$ respondents, and typically has the same range space as $\boldsymbol{Z}$. The empirical distribution of this parameter is commonly characterized in terms of an Item Response Function (IRF) for $n$ respondents and $k$ items or a Person Response Function (PRF) for the $\mathrm{i}^{\text {th }}$ respondent for $k$ items (Meijer \& Sijtsma, 2001). A PRF is a subset of an IRF. The last two parameters represent behavioral responses of items that help characterize a latent trait. The parameter $\boldsymbol{B}$ represents the $l \times k$ vector of difficulty values or observed occurrences of the $k$ items. The parameter $\boldsymbol{A}$ denotes the $l \times k$ vector of discrimination values or differential contribution of the $k$ items.

In this study, the latent trait parameter $\boldsymbol{Z}$ was categorized as the TransfusionRelated Adverse Event (AE). The parameter AE Predisposition ( $\boldsymbol{\theta}$ ) was a measure of a
patient's predisposition to experiencing a particular type of transfusion-related AE. The parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ measured the discrimination and occurrence, respectively, of each type of AE.

## Two-Parameter Logistic IRT Model

Three versions of the two-parameter logistic IRT model were used to estimate the AE Predisposition $(\boldsymbol{\theta})$, discrimination, and difficulty parameters for $k=9$ types of AEs. In that null scores were anticipated and perfect scores were possible for the AE items, a Bayesian framework was used for parameter estimation. Figure 3 (p.30) is presented in order to transition traditional IRT terminology into the modeling of AEs reported for industry clinical trials.

This figure contains two IRFs, each representing a single AE type for 10 fictitious patients. The discrimination and difficulty parameter were -0.5 and -2.0 , respectively, for Curve 1. This curve may be anticipated when patients are relatively homogeneous in relevant baseline parameters. For Curve 2, the discrimination and difficulty parameter were +2.0 and +0.5 , respectively. This curve may be expected when patient populations are relatively heterogeneous in relevant baseline parameters.

With respect to the interpretation of a discrimination parameter, if it is less than zero (Curve 1, Figure 3), the IRF will not be monotonic. This means that high ability persons (large $\boldsymbol{\theta}$ ) will have a smaller probability of correct response than low ability persons (small $\boldsymbol{\theta}$ ). In terms of AEs, this finding would indicate that patients who are most predisposed to a particular type of AE will have lower probabilities of predisposition than
patients who are least predisposed. Basically, negative discrimination values would invert the population risk to a particular type of AE .


Figure 3. Example Two-Parameter Logistic IRT Curves

When a discrimination parameter is between zero and 0.75 , respondents may have comparable probabilities of correct response. In some instances, these items may be viewed as poorly fitting items and removed from the instrument so that model fit is improved (de Gruijter, 2004). In terms of AEs, such a discrimination parameter may indicate that persons in a studied population have a comparable predisposition to a particular type of AE. This is not necessarily a negative finding in an ICT. Furthermore, regulatory agencies require that AE analyses be based on all reported data (ICH E9, 1998).

Last, when a discrimination parameter exceeds 0.75 (Figure 3, Curve 2), the IRF may provide valuable information about instrument quality over the range space of $\boldsymbol{\theta}$. In terms of this study, such a finding may mean that the IRT model is able to differentiate the predisposition of patients experiencing particular types of AEs. It is well known that patients may not be equally predisposed to AEs (Steyerberg et al., 2010), and an IRT model may be a viable solution for identifying patients that are at-risk for experiencing a certain type of AE.

Difficulty is the last parameter to be estimated with an IRT model. If this parameter is negative (Figure 3, Curve 1), the probability of correct response may exceed 0.5 for the majority of study respondents. In relation to this study, many patients having a high predisposition to an AE may be indicative of an at-risk population and/or an unsafe pharmaceutical product. If the difficulty parameter is positive (Figure 3, Curve 2), the probability of correct response may fall below 0.5 for the majority of the respondents (Rudner, 1998). This finding would correlate with a product approved for market use, where the overall patient predisposition to AEs is low for a relatively safe product.

## IRT Model Assumptions

After the estimation of the IRT model parameters, it must be determined whether these results are appropriate for evaluating an instrument. Three key assumptions were pertinent to this evaluation. The first assumption is concerned with the number of latent traits assumed as compared to those that are actually present in a data structure. If an IRT model assumes that one latent trait exists (i.e., unidimensionality) in a data structure and it actually contains two or more traits (i.e., multidimensionality), the resulting parameter
estimates will be biased and should not be used for evaluating an instrument. The second assumption is concerned with whether an IRF is monotonic. Such functions enable respondents with high ability or proficiency to have a higher probability of a correct response on an item than low ability persons. The last assumption, local independence, requires that an instrument item does not affect the probability of correct response to another item; that is, the items are independent.

Dimensionality of Latent Space. This assumption is concerned with identifying the number of dimensions or latent traits in a data structure (Hambleton \& Cook, 1977). This study assumed that the latent trait Transfusion-Related Adverse Event is unidimensional for investigating $k=9$ types of AEs that may occur from the transfusion of blood products such as platelets, whole blood, and plasma. According to Battles, Kaplan, Van der Schaaf, and Shea (1998), the type of AEs reported for a given population, treatment regimen, and clinical indication is typically restricted to the context in which these events occur. This study assumed a single context. If a data structure consisted of two or more contexts, it would be anticipated that the number of Transfusion-Related Adverse Event latent traits would be a function of the number of contexts present. Additional information is provided by Stout (1987), Finch and Habing (2007), and Nandakumar and Stout (1993) on dimensionality for IRT models.

Monotonicity. A CDF is monotonic if it exhibits a non-decreasing form over an $x$-plane (Bartolucci \& Forcina, 2005; Hogg \& Craig, 1995). The logistic CDF for the twoparameter IRT model was employed in this study. This CDF has a known form, and is monotonic (i.e., non-decreasing) over zero to one (Hogg \& Craig, 1995).

Strong Form of Local Independence. The last assumption, conditional upon the respondent's latent trait, requires that "an examinee's performance on one item does not affect his or her performance on other items in the test" (Hambleton \& Cook, 1977, pp. 77-78). This assumption is applicable to ICTs where a patient experiences multiple types of AEs. If the occurrence of one type of AE affects the occurrence of a second AE, these AEs would be dependent, and the strong form of local independence assumption would be violated. For most ICT settings, this is not the case. AEs are typically independent of one another (Grattagliano, Portincasa, Mastronardi, Palmieri, \& Palasciano, 2005; Iannelli, 2010; Laake \& Røttingen, 2001).

## Mathematical Statistics

This study used three forms of the Bayesian two-parameter logistic IRT model to characterize the latent trait Transfusion-Related Adverse Event. Characterization was performed with the AE Predisposition ( $\boldsymbol{\theta})$, discrimination, and difficulty parameters for each of these models. In order to understand how these parameters were estimated, a brief background on mathematical statistics is provided.

This field of statistics is composed of three theoretical frameworks which fall under two philosophical paradigms. The parametric and nonparametric frameworks fall under the classical or frequentist paradigm (Bernardo, 2003). The Bayesian framework falls under the Bayesian paradigm (Bandyopadhyay \& Forster, 2010). These two paradigms differ in structure but attempt to answer the same objectives

Under the classical paradigm, the population parameter $\psi$ is fixed and no distribution is assumed for this parameter (Bartoszyński \& Niewiadomska-Bugaj, 1996).

As a result, prior information can only be used to identify methods appropriate for analyzing data, and determining sample size requirements for substantiating a priori objectives (Lee, 1997). Under the Bayesian paradigm, the population parameter $\psi$ is treated as a random variable (Lynch, 2007) that can be modeled with a "prior" PDF (U.S. Department of Health and Human Services, 2010b). As a result, prior study information can be combined with prospective data to derive a posterior PDF. Hypothesis testing would then be based on estimators from this updated density. Common to both paradigms, statistical inference and power analysis are readily applicable.

## Bayesian Theory

Bayesian statistical inference and power functions are based on posterior densities that are "updated" with prior information (Hogg, McKean, \& Craig, 2005). A combination of univariate and joint densities are needed to derive a posterior PDF. For this derivation, the prior PDF with form $\Psi \sim \mathrm{h}(\psi)$ is to be combined with the likelihood function of the expected data $\mathbf{x} \mid \psi \sim \mathrm{f}(\mathbf{x} \mid \psi)$ and $\psi \in \Psi$. This cross-product results in a joint PDF, for conditionally independent $\mathrm{x}_{\mathrm{i}}$, with form

$$
\begin{equation*}
\mathrm{g}(\mathbf{x}, \psi)=\mathrm{L}(\psi ; \mathbf{x}) \mathrm{h}(\psi) \tag{Eq.2}
\end{equation*}
$$

where $\mathrm{L}(\psi ; \mathbf{x})=\mathrm{f}(\mathbf{x} \mid \psi)=\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}\left(\mathrm{x}_{\mathrm{i}} \mid \psi\right)$ is the likelihood function. The posterior PDF can then be derived as

$$
\begin{equation*}
\mathrm{g}(\psi \mid \mathbf{x})=\frac{\mathrm{L}(\mathbf{x} \mid \psi) \mathrm{h}(\psi)}{\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{~g}(\mathbf{x}, \psi) \mathrm{d} \psi} \tag{Eq.3}
\end{equation*}
$$

$g(\mathbf{x})$ is the unconditional PDF of $x$ with the assumption $g(\mathbf{x}, \psi)=0$ for $\psi \notin[a, b]$.
Example. An example was provided to illustrate the derivation of a posterior PDF and its parameters. This result was then compared to a comparable parametric approach.

Let $\mathbf{X}\left|\theta=X_{1}, \ldots, X_{n}\right| \theta \sim \mathrm{f}(\mathrm{x} \mid \theta)$ be Bernoulli responses $\mathrm{X}_{\mathrm{i}}=\{0,1\}$ where $\mathrm{P}\left(\mathrm{X}_{\mathrm{i}}=1\right)=\theta$ and $\mathrm{P}\left(\mathrm{X}_{\mathrm{i}}=0\right)=1-\theta$. The beta PDF

$$
f(x ; a, b)= \begin{cases}\frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} x^{a-1}(1-x)^{b-1} & 0<x<1 \\ 0 & \text { otherwise }\end{cases}
$$

is the "conjugate prior" (U.S. Department of Health and Human Services, 2010b) of this Bernoulli density. The cross-product of the likelihood function of the Bernoulli PDF and the beta prior results in the joint PDF

$$
\begin{aligned}
g(\mathbf{x} ; \theta) & =\prod_{i=1}^{n} \theta^{x_{i}}(1-\theta)^{1-x_{i}} \cdot \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \\
& =\theta^{\sum_{i=1}^{n} x_{i}}(1-\theta)^{n-\sum_{i=1}^{n} x_{i}} \cdot \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \\
& =\frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \theta^{a+s-1}(1-\theta)^{b+n-s-1}
\end{aligned}
$$

where $s=\sum_{i=1}^{n} X_{i}$. The marginal distribution of this joint density then equals

$$
\begin{aligned}
\mathrm{g}(\mathbf{x}) & =\int_{0}^{1} \mathrm{~g}(\mathbf{x} ; \theta) \mathrm{d} \theta=\int_{0}^{1} \frac{\Gamma(\mathrm{a}+\mathrm{b})}{\Gamma(\mathrm{a}) \Gamma(\mathrm{b})} \theta^{\mathrm{a}+\mathrm{s}-1}(1-\theta)^{\mathrm{b}+\mathrm{n}-\mathrm{s}-1} \mathrm{~d} \theta \\
& =\frac{\Gamma(\mathrm{a}+\mathrm{b})}{\Gamma(\mathrm{a}) \Gamma(\mathrm{b})} \cdot \frac{\Gamma(\mathrm{a}+\mathrm{s}) \Gamma(\mathrm{b}+\mathrm{n}-\mathrm{s})}{\Gamma(\mathrm{a}+\mathrm{b}+\mathrm{n})} .
\end{aligned}
$$

$$
\int_{0}^{1} \frac{\Gamma(\mathrm{a}+\mathrm{b}+\mathrm{n})}{\Gamma(\mathrm{a}+\mathrm{s}) \Gamma(\mathrm{b}+\mathrm{n}-\mathrm{s})} \theta^{\mathrm{a}+\mathrm{s}-1}(1-\theta)^{\mathrm{b}+\mathrm{n}-\mathrm{s}-1} \mathrm{~d} \theta \quad \text { since }[\text { this }]=1
$$

The posterior PDF then equals

$$
\begin{aligned}
\mathrm{g}(\theta \mid \mathbf{x}) & =\frac{\mathrm{g}(\mathbf{x} ; \theta)}{\mathrm{g}(\mathbf{x})}=\frac{\frac{\Gamma(\mathrm{a}+\mathrm{b})}{\Gamma(\mathrm{a}) \Gamma(\mathrm{b})} \theta^{\mathrm{a}+\mathrm{s}-1}(1-\theta)^{\mathrm{b}+\mathrm{n}-\mathrm{s}-1}}{\frac{\Gamma(\mathrm{a}+\mathrm{b})}{\Gamma(\mathrm{a}) \Gamma(\mathrm{b})} \cdot \frac{\Gamma(\mathrm{a}+\mathrm{s}) \Gamma(\mathrm{b}+\mathrm{n}-\mathrm{s})}{\Gamma(\mathrm{a}+\mathrm{b}+\mathrm{n})}} \\
& =\frac{\theta^{\mathrm{a}+\mathrm{s}-1}(1-\theta)^{\mathrm{b}+\mathrm{n}-\mathrm{s}-1}}{\frac{\Gamma(\mathrm{a}+\mathrm{s}) \Gamma(\mathrm{b}+\mathrm{n}-\mathrm{s})}{\Gamma(\mathrm{a}+\mathrm{b}+\mathrm{n})}} \\
& =\frac{\Gamma(\mathrm{a}+\mathrm{b}+\mathrm{n})}{\Gamma(\mathrm{a}+\mathrm{s}) \Gamma(\mathrm{b}+\mathrm{n}-\mathrm{s})} \theta^{\mathrm{a}+\mathrm{s}-1}(1-\theta)^{\mathrm{b}+\mathrm{n}-\mathrm{s}-1} \sim \operatorname{Beta}(a+s, b+n-\mathrm{s})
\end{aligned}
$$

Using this posterior PDF, a posterior probability interval (PPI) or a credibility interval, the Bayesian counterpart to a confidence interval (CI), can be derived. This PPI can then be used, for example, to present the precision of the rate of AEs for a medical treatment (e.g., $0.02 \pm 0.005$ ).

A Wald form of this PPI was based on the first and second moments of the PDF $\operatorname{Beta}(\mathrm{a}+\mathrm{s}, \mathrm{b}+\mathrm{n}-\mathrm{s})$. The conditional mean of this PDF (Hogg \& Craig, 1995) was derived as $E[\theta \mid X=\mathbf{x}]=\frac{a+s}{(a+s)+(b+n-s)}=\frac{a+s}{a+b+n}$, where the scale and location parameter are denoted by $a$ and $b$, respectively. The conditional variance of this function (Hogg \& Craig, 1995) was then derived as

$$
\mathrm{V}[\theta \mid \mathrm{X}=\mathbf{x}]=\frac{(\mathrm{a}+\mathrm{s})(\mathrm{b}+\mathrm{n}-\mathrm{s})}{(\mathrm{a}+\mathrm{b}+\mathrm{n})^{2}(\mathrm{a}+\mathrm{b}+\mathrm{n}+1)}
$$

The resulting PPI was derived as

$$
\frac{a+s}{a+b+n} \pm t_{1-\alpha / 2, n-1} \sqrt{\left(\frac{(a+s)(b+n-s)}{(a+b+n)^{2}(a+b+n+1)}\right) / n}
$$

Continuing with this example, the number of AE occurrences $s$ was set to 10 for $n=100$ patients. The resulting $95 \%$ (two-sided) PPI was $(0.130,0.143)$ for a normal prior density (i.e., $\operatorname{Beta}(5,5))$ and $(0.109,0.121)$ for the prior density $\operatorname{Beta}(3,10)$. Hoehler (1995) developed a classical method for determining sample size requirements based on the respective lower and upper Beta confidence limits $1-\operatorname{Beta}(\mathrm{n}-\mathrm{s}+1, \mathrm{~s})$ and $\operatorname{Beta}(s+1, n-s)$. For $s=10$ and $n=100$, the resulting $95 \%$ (two-sided) CI was ( 0.049 , $0.176)$.

The margins of error for the Bayesian normal and beta priors were 0.0065 and 0.0060 , respectively. The margin of error 0.0635 was found for the classical method. These results demonstrated that the CI is less precise than the PPI for both assumed priors. Furthermore, these results demonstrated the impact of different prior densities on estimation. The prior density that should be used for an analysis should be guided with an understanding of the fit of the prior. If this density is carefully and accurately chosen (Swaminathan et al., 2003) and sample size estimation is sufficiently performed, the obtained precision of the estimate should be trustworthy. This logic is not restricted to one-sample designs, and can be readily applied to more sophisticated statistical approaches such as the IRT model.

## Bayesian Item Response Theory Models

Three different forms of the Bayesian two-parameter logistic IRT model were used to estimate the latent trait parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination, and difficulty. These IRT models can be differentiated as traditional and exchangeable. The traditional model assumed that the latent trait parameters were normal. The exchangeable models had a comparable structure, but differed from the traditional approach in an important way. The exchangeable (EX) and mixed-exchangeable (MEX) models assumed that the distribution of the discrimination parameter was not normal. The MEX model built upon the traditional and EX models by adjusting estimation on outlying discrimination values. This study sought to determine which of these IRT models was most efficient in modeling rare binomial events that possess null data structures. This study characterized this type of data as AEs experienced from the transfusion of blood products such as platelets, whole blood, and plasma.

The general form of the Bayesian IRT model to be utilized was

$$
\begin{equation*}
g\left(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{A}, \mu_{\mathrm{a}}, \mathrm{~s}_{\mathrm{a}}, \mathbf{B} \mid \mathbf{Y}\right) \propto \prod_{\mathrm{i}=1}^{\mathrm{n}} \prod_{\mathrm{j}=1}^{\mathrm{k}}\left[\phi\left(\mathrm{Z}_{\mathrm{ij}}\right) \mathrm{I}\left(\mathrm{Z}_{\mathrm{ij}}, \mathrm{y}_{\mathrm{ij}}\right)\right] \prod_{\mathrm{i}=1}^{\mathrm{n}} \phi\left(\theta_{\mathrm{i}}\right) \prod_{\mathrm{j}=1}^{\mathrm{k}} \phi\left(\mathrm{~b}_{\mathrm{j}}\right) \pi_{0}(\cdot) \tag{Eq.4}
\end{equation*}
$$

$\mathbf{Y}$ is an $n \times k$ matrix of observed occurrences of $k$ AE types for $n$ patients. The values $n$ and $k$ do not need to be a priori unless a study is powered specifically to investigate them. For each AE type $k$, a response of one indicated that patient $i$ experienced a particular type of AE, and a response of zero indicated otherwise. $\mathbf{Z}$, an $n \times k$ matrix of values that represented the latent trait Transfusion-Related Adverse Event, was modeled with the normal conditional posterior PDF

$$
\begin{equation*}
\mathrm{Z}_{\mathrm{ij}} \sim \mathrm{~N}\left(\mu=\mathrm{a}_{\mathrm{j}} \theta_{\mathrm{i}}-\mathrm{b}_{\mathrm{j}}, \sigma^{2}=1 \mid \mathrm{a}_{\mathrm{j}}^{(\mathrm{i})}, \mathrm{b}_{\mathrm{j}}^{(\mathrm{i})}, \theta_{\mathrm{i}}^{(\mathrm{i})}\right) \tag{Eq.5}
\end{equation*}
$$

The superscript (i) denoted the starting or initial value of the AE Predisposition ( $\boldsymbol{\theta}$ ), discrimination, and difficulty parameters. AE Predisposition ( $\boldsymbol{\theta})$, an $n \times 1$ vector of patient predisposition values to experiencing types of AEs, was modeled with the normal conditional PDF

$$
\begin{equation*}
\theta_{\mathrm{i}} \sim N\left(\mu=\frac{\sum_{j=1}^{\mathrm{k}} \mathrm{a}_{\mathrm{j}}\left(\mathrm{Z}_{\mathrm{ij}}+\mathrm{b}_{\mathrm{j}}\right)}{\sum_{\mathrm{j}=1}^{\mathrm{k}}\left(\mathrm{a}_{\mathrm{j}}^{2}+1\right)}, \left.\sigma^{2}=\frac{1}{\sum_{\mathrm{j}=1}^{\mathrm{k}}\left(\mathrm{a}_{\mathrm{j}}^{2}+1\right)} \right\rvert\, \mathrm{a}_{\mathrm{j}}^{(\mathrm{i})}, \mathrm{b}_{\mathrm{j}}^{(\mathrm{i})}, \theta_{\mathrm{i}}^{(\mathrm{i})}, \mathrm{Z}_{\mathrm{ij}}\right) \tag{Eq.6}
\end{equation*}
$$

$\boldsymbol{B}$ was an $l \times k$ vector of difficulty values or observed occurrences of each AE item. $\boldsymbol{A}$ was an $l \times k$ vector of discrimination values that denoted the differential contribution of each AE item to the latent trait. $\boldsymbol{A}$ and $\boldsymbol{B}$ were modeled with the multivariate normal PDF with different forms of $\mu_{a}$ and $s_{a}$, given $\boldsymbol{A}^{(i)}, \boldsymbol{B}^{(i)}, \boldsymbol{Z}$, and $\boldsymbol{\theta}$.

$$
\left[\begin{array}{ll}
\mathbf{A} & \mathbf{B}
\end{array}\right] \sim M N\left(\boldsymbol{\mu}=\left[\mathbf{X}^{\prime} \mathbf{X}+\boldsymbol{\Sigma}_{0}^{-1}\right]^{-1}\left[\begin{array}{l}
\left.\left.\mathbf{X}^{\prime} \mathbf{Z}+\left[\begin{array}{ll}
\mu_{a} & 0
\end{array}\right] \boldsymbol{\Sigma}_{0}^{-1}\right], \boldsymbol{\sigma}^{2}=\left[\mathbf{X}^{\prime} \mathbf{X}+\boldsymbol{\Sigma}_{0}^{-1}\right]\right) \tag{Eq.7}
\end{array}\right.\right.
$$

$\boldsymbol{\Sigma}_{0}^{-1}=\left[\begin{array}{cc}\mathrm{s}_{\mathrm{a}}^{2} & 0 \\ 0 & \mathrm{~s}_{\mathrm{b}}^{2}\end{array}\right]$ and $\boldsymbol{X}$ was the covariate vector $\left(\theta_{\mathrm{i}},-1\right)$. The mean $\left(\mu_{\mathrm{a}}\right)$ and variance $\left(s_{a}^{2}\right)$ of the discrimination parameter were estimated differently for the three IRT models under investigation.

The hierarchical hyperparameters (Lynch, 2007) $\mu_{\mathrm{a}}$ and $\mathrm{s}_{\mathrm{a}}^{2}$ were set to 0 and 1 , respectively, for the two-parameter logistic IRT model or 2-PL (Johnson \& Albert, 1999).

The two-parameter exchangeable logistic IRT model or 2-PL EX (Johnson \& Albert, 1999) used Equation 8 for modeling $\mu_{a}$ and $\mathrm{s}_{\mathrm{a}}^{2}$ :

$$
\begin{equation*}
\pi_{0}\left(\mathbf{A}, \mu_{a}, s_{a}^{2}\right)=\left(s_{a}^{2}\right)^{-v_{1}-1} e^{-v_{2} / s_{a}^{2}} \prod_{j=1}^{k} \phi\left(a_{j} ; \mu_{a}, s_{a}^{2}\right) \tag{Eq.8}
\end{equation*}
$$

Johnson and Albert (1999) recommend $v_{1}=v_{2}=1$. Using the posterior joint PDF comprising this prior PDF the mean

$$
\mu_{\mathrm{a}}=\frac{1}{\mathrm{k}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{a}_{\mathrm{j}}
$$

was first updated with a uniform prior. The variance $s_{a}^{2}$ was then updated by the inverse gamma PDF

$$
\Gamma^{-1} \sim\left(\alpha=\frac{\mathrm{k}}{2}+\mathrm{v}_{1}, \beta=\mathrm{v}_{2}+\frac{1}{2} \sum_{\mathrm{j}=1}^{\mathrm{k}}\left(\mathrm{a}_{\mathrm{j}}-\mu_{\mathrm{a}}\right)^{2}\right)
$$

where $\alpha$ and $\beta$ denoted the scale and location parameter, respectively. The two-parameter mixed-exchangeable logistic IRT model or 2-PL MEX (Toribio, 2006) assumed the conditional posterior PDF of the discrimination parameter was a mixture of two normal density functions in the prior density

$$
\begin{align*}
& \pi_{0}\left(\mathbf{A}, \gamma, \mu_{a}, s_{a}^{2}\right)=\left(s_{a}^{2}\right)^{-v_{1}-1} \\
& \quad \bullet \exp \left\{-\frac{v_{2}}{s_{a}^{2}}\right\} \prod_{j=1}^{k}\left[p \cdot \phi\left(\mu_{a}, s_{a}^{2}\right) \cdot \mathrm{I}\left(\gamma_{j}=0\right)+(1-p) \phi\left(\mu_{a}, K^{2} s_{a}^{2}\right) \cdot \mathrm{I}\left(\gamma_{j}=1\right)\right] \tag{Eq.9}
\end{align*}
$$

for

$$
\begin{equation*}
a_{j} \sim p \cdot \phi\left(\mu_{a}, s_{a}^{2}\right)+(1-p) \cdot \phi\left(\mu_{a}, K^{2} s_{a}^{2}\right) \tag{Eq.10}
\end{equation*}
$$

$\phi$ represented a normal PDF with mean $\mu$ and variance $\sigma^{2} \cdot \gamma$ was a Bernoulli latent trait variable that indicated whether a discrimination value was outlying $\left(\gamma_{\mathrm{j}}=1\right)$ with probability $1-\mathrm{p}$ or not outlying $\left(\gamma_{\mathrm{j}}=0\right)$ with probability p . This probability was estimated as

$$
\begin{equation*}
\mathrm{p}^{(\mathrm{t})}=\frac{\mathrm{p}^{(\mathrm{t}-1)} \phi\left(\mathrm{a}_{\mathrm{j}} ; \mu_{\mathrm{a}}, \mathrm{~s}_{\mathrm{a}}\right)}{\mathrm{p}^{(\mathrm{t}-1)} \phi\left(\mathrm{a}_{\mathrm{j}} ; \mu_{\mathrm{a}}, \mathrm{~s}_{\mathrm{a}}\right)+\left(1-\mathrm{p}^{(\mathrm{t}-1)}\right) \phi\left(\mathrm{a}_{\mathrm{j}} ; \mu_{\mathrm{a}}, \mathrm{Ks}_{\mathrm{a}}\right)} \tag{Eq.11}
\end{equation*}
$$

$\mathrm{p}^{(\mathrm{t}+1)}$ was simulated with the beta PDF with the scale and location parameters set to $n_{1}+1$ and $n_{2}+1$, respectively. $n_{2}$ and $n_{l}$ represented the number of discrimination values $\left(a_{j}\right)$ that were and were not outlying, respectively. If the simulated value was greater than a prespecified value (e.g., 0.90), then the discrimination value was categorized as outlying. For this IRT model $K$ was fixed to a constant of 3 . A uniform prior was assumed for the hyperparameter $\mu_{\mathrm{a}}$ and was updated using the normal PDF with mean

$$
\begin{equation*}
\mu=\frac{\frac{\mathrm{n}_{1}}{\mathrm{~s}_{\mathrm{a}}^{2}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{a}_{\mathrm{j}}^{(\mathrm{t})} \mathrm{I}\left(\gamma_{\mathrm{j}}=0\right)+\frac{\mathrm{n}_{2}}{\mathrm{~K}^{2} \mathrm{~s}_{\mathrm{a}}^{2}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{a}_{\mathrm{j}}^{(\mathrm{t})} \mathrm{I}\left(\gamma_{\mathrm{j}}=1\right)}{\frac{\mathrm{n}_{1}}{\mathrm{~s}_{\mathrm{a}}^{2}}+\frac{\mathrm{n}_{2}}{\mathrm{~K}^{2} \mathrm{~s}_{\mathrm{a}}^{2}}} \tag{Eq.12}
\end{equation*}
$$

An inverse gamma PDF was assumed on the hyperparameter $s_{a}^{2}$ and was updated with variance

$$
\begin{equation*}
\sigma^{2}=\left(\frac{\mathrm{n}_{1}}{\mathrm{~s}_{\mathrm{a}}^{2}}+\frac{\mathrm{n}_{2}}{\mathrm{~K}^{2} \mathrm{~s}_{\mathrm{a}}^{2}}\right)^{-2} \tag{Eq.13}
\end{equation*}
$$

The form of this gamma density had form $\Gamma^{-1} \sim$

$$
\begin{equation*}
\left(v_{1}^{*}=v_{1}+k / 2, v_{2}^{*}=v_{2}+\frac{1}{2} \sum_{j:\left(\gamma_{j}=0\right)}\left(a_{j}-\mu_{a}\right)^{2}+\frac{1}{2 K^{2}} \sum_{j:\left(\gamma_{j}=1\right.}\left(a_{j}-\mu_{a}\right)^{2}\right) \tag{Eq.14}
\end{equation*}
$$

## Gibbs Sampling

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) procedure (Patz \& Junker, 1999a) for simulating parameter estimates from computationally complex joint PDFs (Patz \& Junker, 1999b) such as Equation 4 (p. 38). The inverse cumulative distribution function (CDF) transformation method was used to simulate conditional densities of the joint PDFs so that model parameters could be estimated. This procedure uses the inverse CDF (Johnson \& Liu, 2000) of marginal densities and/or conditional densities to generate random values for estimation.

In order for parameters to be appropriate for use, three estimation assumptions must be reasonably met. First, the mean of each parameter must "converge to the chain's stationary distribution" (Patz \& Junker, 1999b, p. 150). The period prior to convergence is referred to as the burn-in period (Eaves et al., 2005). The stationary state can be visualized as stable means over $i$ iterations of the $x$-plane of the Gibbs sampler. The second assumption requires a "thoroughly mixed" density. A joint density is mixed if the parameter's standard deviation is stable over $i$ iterations of the $x$-plane of the Gibbs sampler. The last assumption, "thinning the chain" (Lynch, 2007, p. 147), pertains to reliability of statistical inference. If the parameter estimates are autocorrelated (May, 2006), the resulting estimators (e.g., mean and variance) may not be appropriate for
decision-making purposes. A correctly designed systematic random sample (Lohr, 1999) for removing serial autocorrelation from parameter estimates may avoid this issue.

To illustrate how a Gibbs sampling algorithms works, a detailed example was provided. Let the joint PDF equal the cross-product of the gamma density $\mathrm{Y} \sim \Gamma(1,1)$, $\mathrm{a}, \mathrm{b}>0$, and the conditional PDF

$$
f(x \mid y)=\left\{\begin{array}{cc}
\mathrm{e}^{-(x-y)}, & 0<y<x<\infty  \tag{Eq.15}\\
0, & \text { otherwise }
\end{array}\right.
$$

The resulting joint PDF was

$$
\begin{equation*}
f(x, y)=f(y) f(x \mid y)=e^{-y} e^{-(x-y)} \tag{Eq.16}
\end{equation*}
$$

Using the inverse CDF transformation method, the CDF of $\mathrm{Y} \sim \Gamma(1,1)$, $\mathrm{a}, \mathrm{b}>0$, was

$$
F(y)=\int_{0}^{y} e^{-t} d t=1-e^{-y}
$$

Next, this CDF was set equal to $u$ and solved for $y$.

$$
\begin{align*}
& \mathrm{u}=1-\mathrm{e}^{-\mathrm{y}} \\
& \Rightarrow \mathrm{e}^{-\mathrm{y}}=1-\mathrm{u} \\
& \Rightarrow \log _{\mathrm{e}}\left(\mathrm{e}^{-\mathrm{y}}\right)=\log _{\mathrm{e}}(1-\mathrm{u}) \\
& \Rightarrow-\mathrm{y}=\log _{\mathrm{e}}(1-\mathrm{u}) \\
& \Rightarrow \mathrm{y}^{*}=-\log _{\mathrm{e}}(1-\mathrm{u})=\mathrm{F}_{\mathrm{Y}}^{-1}(\mathrm{u}), \text { where } \mathrm{u} \sim \mathrm{U}(0,1) \tag{Eq.17}
\end{align*}
$$

$\mathrm{U}(0,1)$ represents the uniform PDF with range space 0,1 . Continuing with this example, the CDF of $f(x \mid y)$ was derived as

$$
F(x \mid y)=\int_{y}^{x} e^{-t+y} d t=e^{y} \int_{y}^{x} e^{-t} d t=e^{y}\left(-e^{-x}+e^{-y}\right)=-e^{-x} e^{y}+1
$$

Setting this result equal to $u$ and solving for $x$ we had

$$
\begin{align*}
& \mathrm{u}=-\mathrm{e}^{-\mathrm{x}} \mathrm{e}^{\mathrm{y}}+1 \\
& \Rightarrow \mathrm{u}-1=-\mathrm{e}^{-\mathrm{x}} \mathrm{e}^{\mathrm{y}} \\
& \Rightarrow 1-\mathrm{u}=\mathrm{e}^{-\mathrm{x}} \mathrm{e}^{\mathrm{y}} \\
& \Rightarrow \log _{\mathrm{e}}(1-\mathrm{u})=\log _{\mathrm{e}}\left(\mathrm{e}^{-\mathrm{x}} \mathrm{e}^{\mathrm{y}}\right) \\
& \Rightarrow \log _{\mathrm{e}}(1-\mathrm{u})=-\mathrm{x}+\mathrm{y} \\
& \Rightarrow \log _{\mathrm{e}}(1-\mathrm{u})-\mathrm{y}=-\mathrm{x} \\
& \Rightarrow \mathrm{y} *-\log _{\mathrm{e}}(1-\mathrm{u})=\mathrm{x}=\mathrm{F}_{\mathrm{X} \mid \mathrm{Y}}^{-1}(\mathrm{u}) \tag{Eq.18}
\end{align*}
$$

Equations 17-18 can then be used to simulate values of $Y$ and $X$, respectively, from the joint $\operatorname{PDFf}(x, y)=e^{-y} e^{-(x-y)}$. With these simulated values, estimators (e.g., mean and variance) of $X$ and $Y$ can be computed across the sequence of Gibbs sampling iterations.

A Gibbs sampling algorithm based on this approach was used to estimate the latent trait parameters $\boldsymbol{Z}$, AE Predisposition $(\boldsymbol{\theta}), \boldsymbol{A}$, and $\boldsymbol{B}$ for the 2-PL, 2-PL EX, and 2-PL MEX IRT models. The general form of the density to be simulated was

$$
\begin{equation*}
\mathrm{g}\left(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{A}, \mu_{\mathrm{a}}, \mathrm{~s}_{\mathrm{a}}, \mathbf{B} \mid \mathbf{Y}\right) \tag{Eq.19}
\end{equation*}
$$

The Gibbs sampling algorithm was performed on three contingent conditional PDFs of the posterior joint PDF. With the initial estimates $\boldsymbol{\theta}^{(i)}, \boldsymbol{A}^{(i)}$, and $\boldsymbol{B}^{(i)}$, the latent variable $\boldsymbol{Z}$ was simulated and estimated with the univariate conditional PDF

$$
\begin{equation*}
\mathrm{g}\left(\mathbf{Z} \mid \mathbf{A}^{(\mathrm{i})}, \mathbf{B}^{(\mathrm{i})}, \boldsymbol{\theta}^{(\mathrm{i})}, \mathbf{Y}\right) \tag{Eq.20}
\end{equation*}
$$

With this parameter, AE Predisposition $(\boldsymbol{\theta})$ was then estimated with the univariate conditional PDF

$$
\begin{equation*}
\mathrm{g}\left(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{A}^{(\mathrm{i})}, \mathbf{B}^{(\mathrm{i})}, \mathbf{Y}\right) \tag{Eq.21}
\end{equation*}
$$

With $\boldsymbol{Z}$ and AE Predisposition $(\boldsymbol{\theta})$ known, $\boldsymbol{A}$ and $\boldsymbol{B}$ were then simultaneously estimated with the multivariate conditional PDF

$$
\begin{equation*}
\mathrm{g}(\mathbf{A}, \mathbf{B} \mid \mathbf{Z}, \boldsymbol{\theta}, \mathbf{Y}) \tag{Eq.22}
\end{equation*}
$$

The assumptions of the sequences of Gibbs sampler iterations will be detailed in Chapter III.

## Clinical Objectives

This section discusses the superiority and equivalence study objectives that were investigated with the 2-PL, 2-PL EX, and 2-PL MEX IRT models.

## Superiority

Superiority is utilized when the objective of a study is to demonstrate that a new group (e.g., Treatment A) is better than or different from another group (e.g., Treatment B) in a given response variable (Lange \& Freitag, 2005). Mathematically, this objective is utilized when a minimum acceptable distance between groups is to be statistically
substantiated. If the effect (e.g., mean of the treatment difference: $\mathrm{A}-\mathrm{B}$ ) of a response variable is expected to be positive, the testing framework should be based on the hypothesis set

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{A}-\mathrm{B} \leq 0 \\
& \mathrm{H}_{1}: A-B>0
\end{aligned}
$$

If the $100(1-\alpha / 2) \%$ lower confidence limit on the effect is greater than zero, then Treatment A is statistically better than Treatment B in a response variable. If the effect is expected to be negative, the testing framework should be based on the hypothesis set
$\mathrm{H}_{0}: \mathrm{A}-\mathrm{B} \geq 0$
$\mathrm{H}_{1}: \mathrm{A}-\mathrm{B}<0$

If the $100(1-\alpha / 2) \%$ upper confidence limit is less than zero, Treatment A is statistically superior to Treatment B in a response variable.

## Hypothesis Testing in IRT Application

This section overviews existing hypothesis testing applied to IRT analyses of binomial item responses. The general concept of each type of IRT objective with references is provided. These hypotheses pertain to the substantiation of calibration for instruments and their items, determination of item invariance, evaluation of model assumptions (dimensionality of latent space, monotonicity, and local independence), and assessment of goodness-of-fit. The latter two uses of IRT hypothesis testing are presented in different sections of this dissertation.

Calibration. Kubinger (2005) indicates that instruments and the items comprising these instruments are calibrated when a certain level of precision of the parameter estimates (e.g., $\boldsymbol{\theta}$ ) is achieved. Although the author uses a likelihood ratio test to assess such precision, a more appropriate analysis may be based on the confidence interval (CI) approach. Significance would be achieved if the limits of the CI for a parameter fall within a predetermined margin of error. As correctly specified by Kubinger, this error should be scientifically based and determined by the researcher prior to the collection of data. Goethals (1994) indicates that the margin of error should not exceed $10 \%$ of the historical range. As an example, if the historical range of a latent trait ability parameter is $1.5-2.5$, then the maximum margin of error would then equal $0.1(2.5-1.5) / 2=0.05$. The researcher could then design a study to statistically substantiate this level of precision with a priori power. If the final study results demonstrated that the confidence limits fall within this margin of error, the researcher could then claim that the item is calibrated. Additional information on hypothesis testing for instrument and item calibration is provided by Goethals (1994), Alexandrowicz (2002), and Bjorner, Kosinski, and Ware (2003).

Differential Item Functioning (DIF). When developing an instrument, it may be desired that its use is not restricted to a limited set of respondent characteristics. That is, the researcher would want the items comprising the instrument to be invariant for as many characteristics as possible. In IRT, this phenomenon pertains to the concept of DIF. If respondents do not differ in latent trait parameters, the applicable instrument is invariant on their characteristics. If DIF is present, then subgroups or specific categories
of subgroups have a significantly different probability of success for a given ability level on a latent trait (Teresi, Kleinman, \& Ocepek-Welikson, 2000). In other words, different types of respondents "may react differently to the same question" (Holland \& Thayer, 1986, p. 1). This defeats the creation of an instrument upon which one can broadly rely.

Two general types of DIF or item bias are available in the literature (Mellenbergh, 1982). Uniform DIF is based on the comparison of IRT models. If a significant difference exists between a full IRT model with all items included and a reduced IRT model with one or more items removed from the analysis, then uniform DIF is present. In other words, this type of DIF exists when Item Response Functions, one per subgroup (e.g., men vs. women), are parallel and significantly different (Swaminathan \& Rogers, 1990). Nonuniform DIF is based on the interaction effect and is evaluated by three types of analysis. First, this type of DIF may occur when an ability parameter $(\boldsymbol{\theta})$ and a subgroup parameter $(\boldsymbol{X})$ interact on a discrimination parameter $(\boldsymbol{A})$ for a particular item (Cook et al., 2007). Next, nonuniform DIF may occur when $\boldsymbol{\theta}$ and $\boldsymbol{X}$ interact on a difficulty parameter (B) for the same item (Cook et al., 2007). Last, nonuniform DIF may occur when IRFs are not parallel, and interaction effects may be present for this scenario (Teresi et al., 2000).

Many approaches have been recommended for determining the presence of DIF, but the Cochran-Mantel-Haenszel statistic (CMH) (Glickman, Seal, \& Eisen, 2009) and logistic regression model (Kok, Mellenbergh, \& van der Flier, 1985) appear to be widely utilized in IRT application.

CMH is based on the adjusted odds ratio (OR) for each item $k$ where the stratification variable is the raw score $(\boldsymbol{Y})$. The hypothesis set is therefore

$$
\begin{aligned}
& \mathrm{H}_{0}: \psi_{\text {adjusted }}=1 \\
& \mathrm{H}_{1}: \psi_{\text {adjusted }} \neq 1
\end{aligned}
$$

If the adjusted OR is significantly different from one, then the groups of interest (e.g., males vs. females) may behave differently on a given item. The OR for each level $r$ of the raw score (stratum) is computed as

$$
\mathrm{OR}_{\mathrm{r}}=\frac{\mathrm{n}_{11_{\mathrm{r}}} \mathrm{n}_{00_{\mathrm{r}}}}{\mathrm{n}_{10_{\mathrm{r}}} \mathrm{n}_{01_{\mathrm{r}}}}
$$

For each stratum, $n_{11}$ denotes the frequency of respondents who achieved the raw score $r$ for both groups of interest (e.g., $1=$ males vs. $0=$ females). $n_{10}$ denotes the frequency of male respondents who achieved the raw score $r$ and the number of female respondents who did not achieve the raw score $r . n_{01}$ denotes the frequency of male respondents who did not achieve the raw score $r$ and the number of female respondents who achieved the raw score $r$. $n_{00}$ denotes the frequency of male and female respondents who did not achieve the raw score $r$ (Sinharay, 2005).

The use of the logistic regression model is a different approach for achieving the same goal as CMH. For each item, the response variable is the item response, the fixed effects are the respondent group variable (e.g., males vs. females) and the item score, and the interaction effect between these predictor variables. A significant interaction effect ( $p$-value $<0.10$ ) denotes the presence of DIF between respondent characteristics.

A key advantage of both the CMH and logistic regression model is that they can be extended to DIF for all items on a test instrument. A possible disadvantage to both of these approaches pertains to statistical power. If studies are not powered to sufficiently
detect a relevant magnitude of DIF, then conclusions regarding how various groups compare on a latent trait could be confounded with sample size. Additional information for evaluating DIF is provided by Crane, Gibbons, Narasimhalu, Lai, and Cella (2007), Perkins, Stump, Monahan, and McHorney (2006), Steinberg and Thissen (2006), Froelich and Habing (2008), Lewis (2006), Stone (2003), and Douglas (1996).

## Equivalence

Equivalence, loosely speaking, can be thought of as the antithesis of superiority. When the objective of a study is to demonstrate that one treatment is the same as or comparable to another treatment in a response variable, equivalence is the recommended study objective (Liu, 2005). The equivalence margin ( $\Delta$ ), which will be detailed in the next section, is a measure of the degree of this similarity or comparability (Bauer, 2005; Blackwelder, 2002). Mathematically, this objective is utilized when a maximum acceptable distance between groups is to be statistically substantiated. As opposed using a single one- or two-sided hypothesis to evaluate an objective, equivalence is based on two one-sided hypothesis tests in the set for Treatment A and B

$$
\begin{aligned}
H_{0}: A & -B \leq-\Delta B \text { or } A-B \geq+\Delta B \\
& \Leftrightarrow A-B+\Delta B \leq 0 \text { or } A-B-\Delta B \geq 0 \\
H_{1}:-\Delta & B<A-B<+\Delta B \\
& \Leftrightarrow \omega_{1}: A-B+\Delta B>0 \text { and } \omega_{2}: A-B-\Delta B<0
\end{aligned}
$$

If the $100(1-\alpha / 2) \%$ lower confidence limit on effect $1\left(\omega_{1}\right)$ is greater than zero and the $100(1-\alpha / 2) \%$ upper confidence limit on effect $2\left(\omega_{2}\right)$ is less than zero, then equivalence
is statistically substantiated between the treatments groups on a response variable with $100(1-\alpha) \%$ confidence.

## Equivalence Margin Estimation

The equivalence margin specifies how far apart treatment groups can be before they are no longer scientifically comparable in a given response variable. Margins should be determined only in an a priori manner, and are necessary for correct sample size estimation. Analyses based on such a planned sample size should only use this margin. The study or trial may not be powered appropriately for other margins.

The available approaches or definitions for statistically estimating this margin for non-placebo trials were provided. After an equivalence margin is statistically estimated, it should be scientifically or clinically validated before use (Lange \& Freitag, 2005). This process comprises determining the impact on study conclusions for the chosen equivalence margin. As an extreme example, if one treatment group should be within $10 \%$ of another, the study should not be designed on a margin 5 times larger than the expected difference.

Equivalence Margin 1. The first definition multiplies a historical difference between treatments A and B (Hung, Wang, Tsong, Lawrence, \& O'Neil, 2003) by the constant $k$. This margin is estimated as

$$
\begin{equation*}
\Delta=\mathrm{k}\left(\overline{\mathrm{X}}_{\mathrm{A}}-\overline{\mathrm{X}}_{\mathrm{B}}\right) \tag{Eq.23}
\end{equation*}
$$

The use of this difference must be closely evaluated in relation to historical and prospective study parameters to be employed (e.g., design, population, location) (ICH E10, 2000). The constant $k$ may range from 0 to 1 , where these values denote superiority
and absolute equivalence, respectively, between treatment groups. Although Hauschke (2001) recommends that $k$ can equal $1 / 5,1 / 3$, or $1 / 2$, this constant should be based on the desired effect to be retained. For example, if $k=0.1$, then the study would be designed to demonstrate that at least $90 \%$ of the effect from Treatment B is retained by Treatment A. Ng (2001) indicates that the value of $k$ should be balanced with clinical substantiation and sample size feasibility.

Equivalence Margin 2. The next definition for estimating an equivalence margin multiplies the historical standard deviation of an effect by the constant $k$ (Wiens, 2002).

Equivalence Margin 3. The third definition for estimating an equivalence margin is based on a fraction of Cohen's effect size of the historical data ( $\mathrm{Ng}, 2001$ ). This margin is estimated as

$$
\begin{equation*}
\Delta=\mathrm{k}\left(\frac{\overline{\mathrm{X}}_{\mathrm{A}}-\overline{\mathrm{X}}_{\mathrm{B}}}{\mathrm{~s}_{\mathrm{AB}}}\right) \tag{Eq.25}
\end{equation*}
$$

Equivalence Margin 4. The fourth definition for estimating an equivalence margin is based on the Proportion of Similar Responses (PSR) approach (Heyse \& Stine, 2000). PSR estimates the margin by calculating the area under empirical densities $\mathrm{f}_{\mathrm{i}}(\mathrm{x})$. This margin is estimated as

$$
\begin{equation*}
\operatorname{PSR}=\int_{-\infty}^{\infty} \min \left[\mathrm{f}_{\mathrm{i}}(\mathrm{x})\right] \mathrm{dx} \tag{Eq.24}
\end{equation*}
$$

Equivalence Margin 5. Chow and Shao (2006) developed a definition for estimating an equivalence margin that is based on the standard normal density function and the variability of historical data. For treatments A and B , this margin equals

$$
\begin{equation*}
\Delta=\left(\mathrm{Z}_{1-\alpha}+\mathrm{Z}_{\eta}\right)\left(\frac{\sigma_{\mathrm{B}}^{2}}{\mathrm{n}_{\mathrm{B}}}+\mathrm{c}^{2}\right)^{1 / 2}-\mathrm{Z}_{1-\varepsilon}\left(\frac{\sigma_{\mathrm{A}}^{2}}{\mathrm{n}_{\mathrm{A}}}+\mathrm{c}^{2}\right)^{1 / 2} \tag{Eq.26}
\end{equation*}
$$

where $\boldsymbol{Z}$ denotes a standard normal critical value, $\eta$ is the desired statistical power, $Z_{l-\varepsilon}=$ $Z_{1-\alpha}+Z_{\beta}$, where $\beta$ is the allowed Type II error (e.g., $20 \%$ for a Phase III ICT), and $\mathrm{c}=\min \left\{\frac{\sigma_{\mathrm{B}}}{\mathrm{n}_{\mathrm{B}}^{1 / 2}}, \frac{\sigma_{\mathrm{A}}}{\mathrm{n}_{\mathrm{A}}^{1 / 2}}\right\}$.

Equivalence Margin 6. The last definition, referred to as the $95 \%-95 \%$ approach, for estimating an equivalence margin was developed and recommended by the FDA (U.S. Department of Health and Human Services, 2010a). This method first requires estimating the $97.5 \%$ lower confidence limit of a $95 \%$ (two-sided) CI on the percent difference between treatment groups for the desired variable. With this lower limit of this CI, the equivalence margin is computed as

$$
\begin{equation*}
\mathrm{k}\left(1-\frac{1}{1-\mathrm{LCL}}\right) \tag{Eq.28}
\end{equation*}
$$

The six equivalence margin definitions will be quantitatively compared in Chapter IV for various rates of binomial response variables (Table 30, p. 166). In this study, a single arbitrary equivalence margin $(\Delta=10 \%)$ was used for statistical inference and power analysis application.

## Power Analysis

This section reviews the available literature on power analysis used for IRT and paired binomial application. Some of this literature provides general guidance simply recommending its use to the reader when designing instruments. Other authors provide
general sample size requirements based on Cohen's effect size. Studies that utilized formal power analysis for IRT application are then presented. After background on power analysis in IRT utilization is provided, this section introduces the concept of paired binomial responses. The existing functions for computing asymptotic Type I and II errors are then presented for multiple superiority and equivalence approaches.

Three technical definitions of statistical power $(1-\beta)$ are generally available. The first definition indicates that power is the probability of rejecting the null hypothesis given the alterative hypothesis is true. The second technical definition indicates that power is the probability of rejecting the null hypothesis when it is false. Power has also been formally defined as the probability of rejecting the null hypothesis "for any given set of circumstances, even those corresponding to $\mathrm{H}_{0}$ being true." (Castelloe, 2000, p. 1).

These conditional definitions of power are needed in order to have a universal definition for different types of hypothesis sets used for statistical inference. Composite hypotheses (Neyman, 1977) such as $\mathrm{H}_{0}: \mu_{\mathrm{A}}-\mu_{\mathrm{B}} \geq 0$ versus $\mathrm{H}_{1}: \mu_{\mathrm{A}}-\mu_{\mathrm{B}}<0$ cover the full range of possible treatment differences. The simple hypothesis set (Graves, 1978) $\mathrm{H}_{0}: \mu_{\mathrm{A}}-\mu_{\mathrm{B}}=\mathrm{k}_{0}$ versus $\mathrm{H}_{1}: \mu_{\mathrm{A}}-\mu_{\mathrm{B}}=\mathrm{k}_{1}$ and the hybrid hypothesis set (Jones, 1952) $\mathrm{H}_{0}: \mu_{\mathrm{A}}-\mu_{\mathrm{B}}=0$ versus $\mathrm{H}_{1}: \mu_{\mathrm{A}}-\mu_{\mathrm{B}}<0$ do not cover the full range of treatment differences. As a result, the condition that the alternative hypothesis be true accommodates these three testing scenarios.

Item Response Theory

The Western Michigan University library system, WorldCat, Web of Science, ERIC, and ABI/Inform document databases were searched for all combinations of the terms item response theory, power function, and power analysis. The database results returned a total of nine articles. Five of the articles specified that power analysis was performed, but details for conducting the power analysis were not provided. The remaining articles presented either a formal power function or the details for computing sample size requirements for IRT application.

Eisen, Wilcox, Leff, Schaefer, and Culhane (1999) recorded mental health variables based on the Behavior and Symptom Identification Scale (BASIS-32) for a small sample of severely ill patients. The authors concluded that their small sample size was a limitation of their study with respect to their primary analysis based on Classical Test Theory (Lord \& Novick, 1968). The authors concluded that to overcome this limitation IRT may "create useful, reliable, and valid outcome instruments" (Eisen et al., 1999, p. 15) for studies that have adequate statistical power. May, Cole, Haimson, and Perez-Johnson (2011) investigated achievement levels of state assessment exams. The purpose of their study was to derive parameters for these exams enabling more reliable sample size estimation on the ability parameter $\boldsymbol{\theta}$ for various combinations of $\boldsymbol{A}$ and $\boldsymbol{B}$ for future studies. Steinberg and Thissen (2006) discussed sample size estimation based on group differences in the IRT parameters discrimination and difficulty. Cohen's effect sizes on these parameters were utilized to determine the minimum number of respondents required for detecting DIF between groups on these parameters. Cole, Lin, and Rupnow
(2009) conducted a study to determine minimally important effect sizes for treatment groups based on the Migraine-Specific Quality of Life Questionnaire version 2.1 (MSQ v2.1). Cohen effect sizes, based on the mean of change score differences, were based on three doses of Toperimate versus a placebo for the treatment of migraines. The article recommended that the study results could be used to power future trials for this investigational treatment. Kubinger (2005) used likelihood ratio tests to discuss the relationship between sample size requirements and test calibration for adaptive testing. In order for a test to be calibrated, the parameter estimates from the IRT model must be "accurate" (p. 388). The author further indicated that the definition of accuracy must have "practical relevance" (p.389), and went on to state that the best means of minimizing irrelevant significance tests is to appropriately power a study for the desired effect.

In summary, the articles that discuss sample size recommendations based on IRT application and methodology are general and based on preconceived estimators such as Cohen's effect size. In terms of formal power analysis, such an approach is vague and not recommended. The estimation of effects (numerator of effect size) is always specific to the study objective of interest. Furthermore, effects are computed differently for $k$-sample matched versus $k$-sample independent study designs, and also between superiority and equivalence study objectives. Also, effect sizes are dependent upon the type of estimator (e.g., mean, median, slope) employed for the effect. As a result, sample size requirements are typically specific to the study being conducted.

The remaining IRT articles presented formal power functions or sufficient detail for determining sample size requirements. Sato, Rabinowitz, Gallagher, and Huang (2010) utilized a Difference-in-Difference study design for investigating items on a
mathematics exam. One-, two-, and three-parameter logistic IRT models were used for generating ability estimates for two types of math item sets (linguistically modified (L), original (O)) for three groups (EL, NEP, EP) of students categorized for reading proficiency levels. The hypothesis set for testing the differences between these groups was

$$
\begin{aligned}
& \mathrm{H}_{0}: \bar{\theta}_{\mathrm{EL}_{\mathrm{L}-\mathrm{O}}}-\bar{\theta}_{\mathrm{NEP} \mathrm{~L}_{\mathrm{L}} \mathrm{O}}-\bar{\theta}_{\mathrm{EP}_{\mathrm{L}-\mathrm{O}}}=0 \\
& \mathrm{H}_{1}: \bar{\theta}_{\mathrm{EL}_{\mathrm{L}-\mathrm{O}}}-\bar{\theta}_{\mathrm{NEP} \mathrm{~L}^{-O}}-\bar{\theta}_{\mathrm{EP}_{\mathrm{L}-\mathrm{O}}} \neq 0
\end{aligned}
$$

The authors presented the following formula for determining the minimum detectable effect size

$$
\mathrm{ES}=\left(\frac{\left(\mathrm{Z}_{1-\alpha / 2}+\mathrm{Z}_{1-\beta}\right)^{2}\left(\sum \mathrm{C}_{\mathrm{i}}^{2}\right) \sigma^{2}}{\mathrm{n}_{\mathrm{i}}}\right)^{1 / 2}
$$

for level $\alpha$ and power $1-\beta . C_{i}$ is the contrast coefficient for cell $i, \sigma^{2}$ is the common population (within-cell) variance, and $n_{i}$ is the harmonic mean of the sample sizes in the $k$ cells. The power function can be readily extracted from the above equation by solving for $1-\beta$. With this function, the probability of detecting any effect size can be computed for a desired sample size and confidence level.

Holman, Math, Glas, and de Haan (2003) conducted a study that investigated responses to the health-status surveys $\mathrm{SF}-36, \mathrm{SF}-12$, and $\mathrm{SF}-8$. The two-parameter logistic IRT model was utilized to investigate binary medical outcomes. The primary variable was the ability parameter $(\boldsymbol{\theta})$ that measured a single latent trait. Treatment differences $(\mathrm{A}-\mathrm{B})$ in $\boldsymbol{\theta}$ were tested with the hypothesis set

$$
\begin{aligned}
& \mathrm{H}_{0}: \bar{\theta}_{\mathrm{A}}-\bar{\theta}_{\mathrm{B}}=0 \\
& \mathrm{H}_{1}: \bar{\theta}_{\mathrm{A}}-\bar{\theta}_{\mathrm{B}} \neq 0 .
\end{aligned}
$$

Bootstrap methodology was utilized to simulate the power of achieving the alternative hypothesis for various effect sizes and sample sizes per treatment group.

Stone (2003) also used simulation methodology to investigate the empirical power of the Pearson $\left(\chi^{* 2}\right)$ and Likelihood Ratio $\left(G^{*^{2}}\right)$ goodness-of-fit (GoF) statistics to detect misfitting items. Simulations comprised 100 Monte Carlo studies (MCs) each with 200 bootstraps with 500 to 2,000 respondents per bootstrap. Hypothesis tests were based on differences in the discrimination parameter between instrument items, where this parameter was incremented by 0.3 . This study found that the GoF statistics were unable to detect item misfit for discrimination parameter differences of less than or equal to 0.3 . When this difference was increased to 0.5 for sample sizes exceeding 1,000 , both GoF statistics were able to detect item misfit with sufficient statistical power. Overall, the Likelihood Ratio GoF statistic was found to be "slightly less" (p. 579) powerful than the Pearson chi-square GoF statistic (Stone, 2003).

Last, Lord (1953) presented methodology for statistically identifying a single psychological test from a set of tests that bests discriminates "successful" examinees for a particular trait. Post-hoc power analysis based on Neyman-Pearson hypothesis testing (Bartoszyński \& Niewiadomska-Bugaj, 1996) was employed for this purpose. Hypothesis testing, based on ability scores estimated with IRT, was used for statistical power calculations. The hypothesis set for each psychological test (i) employed was

$$
\begin{aligned}
& \mathrm{H}_{0}: \hat{\theta}_{\mathrm{i}} \leq \theta \\
& \mathrm{H}_{1}: \hat{\theta}_{\mathrm{i}}>\theta
\end{aligned}
$$

where $\boldsymbol{\theta}$ was the true ability score and $\hat{\theta}_{\mathrm{i}}$ was the estimated maximum likelihood estimate of the ability score. This type of analysis resulted in the post-hoc power or probability of each test achieving the desired objective. In Lord's study, the test having the highest power was the one that was best able to discriminate examinees.

## Paired Binomial Endpoints

For this study, analysis for classical and IRT application was based on the difference between correlated variables. This section presents the available power functions for comparing correlated binomial variables or dependent proportions for the $k$ sample matched-paired design. Discussion is limited to those articles that present formal sample size methodology for the objectives superiority (Bishop, Feinberg, \& Holland, 1975; Connett, Smith, \& McHugh, 1987; Connor, 1987; Duffy, 1984; Lachenbruch, 1992; Lachin, 1992; Miettinen, 1968; Mitra, 1958; Schlesselman, 1982) and equivalence (Liu, Hsueh, Hsieh, \& Chen, 2002). Power functions for the multiple-match case or dependent repeated measures design (Dupont, 1988) are not presented in this dissertation. Additionally, power functions for the noninferiority objective, which reflects one side of an equivalence test, are available in Nam (1997), Lu and Bean (1995), Tang, Tang, Chan, and Chan (2002), Lee and Lusher (1991), Chow and Shao (2006), and Tango (1998). Additional authors present sample size requirements based on CIs, but these calculations are typically not a function of statistical power (Lachenbruch \& Lynch, 1998; Pham-Gia
\& Turkkan, 2008). Last, other authors present hypotheses and associated estimators (e.g., mean, variance, CI), but powering a study based on their method(s) was not provided (Agresti \& Min, 2005; Altham, 1971; Koroto, 2009; Lloyd, 1990; May \& Johnson, 1997; Newcombe, 1998; Wacholder \& Weinberg, 1982). Before this literature is presented, a brief discussion on $2 \times 2$ contingency tables is provided.

A study design comprising two treatment groups for comparing AEs can be represented by the $2 \times 2$ contingency format presented in Table 1 . The main-diagonal represents the frequency $(x)$ and proportion $(p)$ of concordant pairs. In thisstudy, $x_{00}$ and $x_{11}$ represented the number of patients that did not and did, respectively, experience a particular type of AE for both treatment groups. The off-diagonal represented the frequency and proportion of discordant pairs. $x_{10}$ represented the number of patients that experienced a particular type of AE on Treatment A but not Treatment B. $x_{01}$ represented the number of patients that experienced a particular type of AE on Treatment B but not Treatment A.

Table 1
$2 \times 2$ Contingency Table for Binomial Events

|  | Event Occurred | Treatment B |  | Marginals |
| :---: | :---: | :---: | :---: | :---: |
|  |  | Result $=1$ | Result $=0$ |  |
|  |  |  |  |  |
| Treatment A | Result $=1$ | $\mathrm{x}_{11} / \mathrm{p}_{11}$ | $\mathrm{x}_{10} / \mathrm{p}_{10}$ | $\mathrm{p}_{1}=\mathrm{p}_{11}+\mathrm{p}_{10}$ |
|  | Result $=0$ | $\mathrm{x}_{01} / \mathrm{p}_{01}$ | $\mathrm{x}_{00} / \mathrm{p}_{00}$ | $\mathrm{q}_{1}$ |
| Marginals |  | $\mathrm{p}_{2}=\mathrm{p}_{11}+\mathrm{p}_{01}$ | $\mathrm{q}_{2}$ | $\Sigma \mathrm{p}_{\mathrm{ij}}=1$ |

## Type I and II Errors for Superiority

This section describes multiple methods for computing type 1 and 2 errors for the paired study design for variables with binomial responses. Methods 1 and 2 are based on the odds ratio (OR) of discordant probabilities, and the remaining methods are based on the difference between these probabilities. All of these methods use various unconditional and conditional approaches to derive asymptotic modifications of the assumed effects (ORs or differences) for McNemar's statistic. In that there is no guidance as to which of these methods are recommended for rare binomial events, this study compared sample size requirements between the existing methods and the IRT approaches derived by the author.

Superiority Method 1. Connett, Smith, and McHugh (1987) developed an unconditional approach for determining sample size requirements based on the OR of discordant probabilities $p_{a b}$ for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \psi=\frac{\mathrm{p}_{10}}{\mathrm{p}_{01}} \geq 1 \\
& \mathrm{H}_{1}: \psi=\frac{\mathrm{p}_{10}}{\mathrm{p}_{01}}<1
\end{aligned}
$$

$p$ denotes the cell proportion for cell $\mathrm{a}, \mathrm{b}(a=$ Treatment $\mathrm{A}, b=$ Treatment B$)$. In an ICT investigating AEs, the property $\mathrm{p}_{10}<\mathrm{p}_{01}$ would be desired. The function used to determine the number of units of analysis (e.g., persons, items) for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\frac{\left[\mathrm{Z}_{\alpha}(\psi+1)^{1 / 2}+\mathrm{Z}_{\beta}\left\{(\psi+1)-(\psi-1)^{2} \mathrm{p}_{01}\right\}^{1 / 2}\right]^{2}}{(\psi-1)^{2} \mathrm{p}_{01}} \tag{Eq.29}
\end{equation*}
$$

where $Z$ is a normal critical value. The power function, algebraically solved from Equation 29, is

$$
\begin{equation*}
1-\Phi\left[\frac{(\psi-1)\left(n p_{01}\right)^{1 / 2}-\mathrm{Z}_{\alpha}(\psi+1)^{1 / 2}}{\left\{(\psi+1)-(\psi-1)^{2} \mathrm{p}_{01}\right\}^{1 / 2}}\right] \tag{Eq.30}
\end{equation*}
$$

$\Phi(\cdot)$ denotes the standard normal quantile, and is computed with the PROBNORM function in SAS. The observed significance level, algebraically solved from Equation 29, is computed with the function

$$
\begin{equation*}
\Phi\left[\frac{(\psi-1)\left(\mathrm{np}_{01}\right)^{1 / 2}-\mathrm{Z}_{\beta}\left\{(\psi+1)-(\psi-1)^{2} \mathrm{p}_{01}\right\}^{1 / 2}}{(\psi+1)^{1 / 2}}\right] \tag{Eq.31}
\end{equation*}
$$

Superiority Method 2. Schlesselman (1982) derived a conditional method for determining sample size requirements based on the OR of discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \psi=\frac{\mathrm{p}_{10}}{\mathrm{p}_{01}} \geq 1 \\
& \mathrm{H}_{1}: \psi=\frac{\mathrm{p}_{10}}{\mathrm{p}_{01}}<1
\end{aligned}
$$

The function used to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\frac{\left[\mathrm{Z}_{\alpha}(\psi+1)+2 \mathrm{Z}_{\beta} \sqrt{\psi}\right]^{2}}{(\psi-1)^{2}(\psi+1) \mathrm{p}_{01}} \tag{Eq.32}
\end{equation*}
$$

The power function, algebraically solved from Equation 32, is

$$
\begin{equation*}
1-\Phi\left[\frac{(\psi-1)\left[\mathrm{n}(\psi+1) \mathrm{p}_{01}\right]^{1 / 2}-\mathrm{Z}_{\alpha}(\psi+1)}{2 \psi^{1 / 2}}\right] \tag{Eq.33}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 32, is computed with the function

$$
\begin{equation*}
\Phi\left[\frac{(\psi-1)\left[(\psi+1) \mathrm{p}_{01}\right]^{1 / 2}-2 \mathrm{Z}_{\beta} \psi^{1 / 2}}{\psi+1}\right] \tag{Eq.34}
\end{equation*}
$$

Superiority Method 3. Connor (1987) derived a method for determining sample size requirements based on the difference in discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq 0 \\
& \mathrm{H}_{1}: \mathrm{p}_{10}-\mathrm{p}_{01}<0
\end{aligned}
$$

The function used to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\frac{\left[\mathrm{Z}_{\alpha} \varphi^{1 / 2}+\mathrm{Z}_{\beta}\left(\varphi-\delta^{2}\right)^{1 / 2}\right]^{2}}{\delta^{2}} \tag{Eq.35}
\end{equation*}
$$

for $\varphi=\mathrm{p}_{10}+\mathrm{p}_{01}$ and $\delta=\mathrm{p}_{10}-\mathrm{p}_{01}$. The power function, algebraically solved from Equation 35, is

$$
\begin{equation*}
1-\Phi\left[\frac{\delta n^{1 / 2}-\mathrm{Z}_{\alpha} \varphi^{1 / 2}}{\left(\varphi-\delta^{2}\right)^{1 / 2}}\right] \tag{Eq.36}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 35, is computed with the function

$$
\begin{equation*}
\Phi\left[\frac{\delta n^{1 / 2}-\mathrm{Z}_{\beta}\left(\varphi-\delta^{2}\right)^{1 / 2}}{\varphi^{1 / 2}}\right] \tag{Eq.37}
\end{equation*}
$$

Superiority Method 4. Lachenbruch (1992) derived a method for determining sample size requirements based on the difference in discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq 0 \\
& \mathrm{H}_{1}: \mathrm{p}_{10}-\mathrm{p}_{01}<0
\end{aligned}
$$

The function used to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\frac{0.25\left(\mathrm{Z}_{\alpha}+\mathrm{Z}_{\beta}\right)^{2}}{(0.5-\mathrm{s})^{2}\left|\mathrm{p}_{+1}+\mathrm{p}_{1+}-2 \mathrm{p}_{00}\right|} \tag{Eq.38}
\end{equation*}
$$

where $\mathrm{p}_{+1}=\frac{\mathrm{x}_{00}+\mathrm{x}_{01}}{\mathrm{n}}, \mathrm{p}_{1+}=\frac{\mathrm{x}_{00}+\mathrm{x}_{10}}{\mathrm{n}}$, and $\mathrm{s}=\frac{\mathrm{p}_{+1}-\mathrm{p}_{00}}{\mathrm{p}_{+1}+\mathrm{p}_{1+}-2 \mathrm{p}_{00}}$. The power function, algebraically solved from Equation 38, is

$$
\begin{equation*}
1-\Phi\left[(0.5-\mathrm{s})\left\{\frac{\mathrm{n}}{0.25}\left|\mathrm{p}_{+1}+\mathrm{p}_{1+}-2 \mathrm{p}_{00}\right|\right\}^{1 / 2}-\mathrm{Z}_{\alpha}\right] \tag{Eq.39}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 39, is computed with the function

$$
\begin{equation*}
\Phi\left[(0.5-\mathrm{s})\left\{\frac{\mathrm{n}}{0.25}\left|\mathrm{p}_{+1}+\mathrm{p}_{1+}-2 \mathrm{p}_{00}\right|\right\}^{1 / 2}-\mathrm{z}_{\beta}\right] \tag{Eq.40}
\end{equation*}
$$

Superiority Method 5. Miettinen (1968) derived a first-order unconditional method for determining sample size requirements based on the difference in discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq 0 \\
& \mathrm{H}_{1}: \mathrm{p}_{10}-\mathrm{p}_{01}<0
\end{aligned}
$$

The formula used to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\left[\frac{\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}+\mathrm{Z}_{\beta} \sqrt{\Sigma^{2} / \mathrm{n}}}{\delta}\right]^{2} \tag{Eq.41}
\end{equation*}
$$

where $\bar{\pi}=\varphi / 2$ and $\Sigma^{2}=\operatorname{Var}[\mathrm{X} \mid \mathrm{E}(\mathrm{M})]=4 \mathrm{np}_{10} \mathrm{p}_{01} / \varphi$ for the expected number of discordant events (M). The power function, algebraically solved from Equation 41, is

$$
\begin{equation*}
1-\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}}{\sqrt{\Sigma^{2} / \mathrm{n}}}\right) \tag{Eq.42}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 41, is computed with the function

$$
\begin{equation*}
\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\beta} \sqrt{\Sigma^{2} / \mathrm{n}}}{\sqrt{2 \bar{\pi}}}\right) \tag{Eq.43}
\end{equation*}
$$

Superiority Method 6. Miettinen (1968) derived a second-order unconditional method function for determining sample size requirements based on the difference in discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq 0 \\
& \mathrm{H}_{1}: \mathrm{p}_{10}-\mathrm{p}_{01}<0
\end{aligned}
$$

The formula to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\left[\frac{\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}+\mathrm{Z}_{\beta} \sqrt{\Sigma^{2 / n}}}{\delta}\right]^{2} \tag{Eq.44}
\end{equation*}
$$

where $\Sigma^{2}=\operatorname{Var}[\mathrm{X} \mid \mathrm{E}(\mathrm{M})]=\frac{\mathrm{n}}{\varphi}\left[\varphi^{2}+\delta^{2}(3+\varphi) / 4\right]$ for the expected number of discordant events (M). The power function, algebraically solved from Equation 44, is

$$
\begin{equation*}
1-\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}}{\sqrt{\Sigma^{2} / \mathrm{n}}}\right) \tag{Eq.45}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 44, is computed with the function

$$
\begin{equation*}
\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\beta} \sqrt{\Sigma^{2} / \mathrm{n}}}{\sqrt{2 \bar{\pi}}}\right) \tag{Eq.46}
\end{equation*}
$$

Superiority Method 7. Bishop, Feinberg, and Holland (1975) derived a multinomial unconditional method for determining sample size requirements based on the difference in discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq 0 \\
& \mathrm{H}_{1}: \mathrm{p}_{10}-\mathrm{p}_{01}<0
\end{aligned}
$$

The formula used to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\left[\frac{\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}+\mathrm{Z}_{\beta} \sqrt{\Sigma^{2} / \mathrm{n}}}{\delta}\right]^{2} \tag{Eq.47}
\end{equation*}
$$

where $\Sigma^{2}=\operatorname{Var}[\mathrm{X} \mid \mathrm{E}(\mathrm{M})]=\mathrm{n}\left(\varphi-\delta^{2}\right)$ for the expected number of discordant events $(M)$. The power function, algebraically solved from Equation 47, is

$$
\begin{equation*}
1-\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}}{\sqrt{\Sigma^{2} / \mathrm{n}}}\right) \tag{Eq.48}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 47, is computed with the function

$$
\begin{equation*}
\Phi\left(\frac{\delta \sqrt{n}-Z_{\beta} \sqrt{\Sigma^{2} / n}}{\sqrt{2 \bar{\pi}}}\right) \tag{Eq.49}
\end{equation*}
$$

Superiority Method 8. Mitra (1958) derived a local unconditional method for determining sample size requirements based on the difference in discordant probabilities for the following superiority hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq 0 \\
& \mathrm{H}_{1}: \mathrm{p}_{10}-\mathrm{p}_{01}<0
\end{aligned}
$$

The formula used to determine the number of units of analysis for this hypothesis set is

$$
\begin{equation*}
\mathrm{n}=\left[\frac{\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}+Z_{\beta} \sqrt{\Sigma^{2} / \mathrm{n}}}{\delta}\right]^{2} \tag{Eq.50}
\end{equation*}
$$

where $\Sigma^{2}=\operatorname{Var}[\mathrm{X} \mid \mathrm{E}(\mathrm{M})]=\mathrm{n} \varphi$ for the expected number of discordant events $(M)$. The power function, algebraically solved from Equation 50, is

$$
\begin{equation*}
1-\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\alpha} \sqrt{2 \bar{\pi}}}{\sqrt{\Sigma^{2} / \mathrm{n}}}\right) \tag{Eq.51}
\end{equation*}
$$

The observed significance level, algebraically solved from Equation 50, is computed with the function

$$
\begin{equation*}
\Phi\left(\frac{\delta \sqrt{\mathrm{n}}-\mathrm{Z}_{\beta} \sqrt{\Sigma^{2} / \mathrm{n}}}{\sqrt{2 \bar{\pi}}}\right) \tag{Eq.52}
\end{equation*}
$$

Type I and II Errors for Equivalence

Equivalence Method 1. Liu, Hsueh, Hsieh, and Chen (2002) derived an asymptotic Wald-based power function based on the difference in discordant probabilities for the following equivalence hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \delta=\mathrm{p}_{10}-\mathrm{p}_{01} \geq \Delta \text { or } \mathrm{p}_{10}-\mathrm{p}_{01} \leq-\Delta \\
& \mathrm{H}_{1}:-\Delta<\mathrm{p}_{10}-\mathrm{p}_{01}<\Delta
\end{aligned}
$$

The power function is

$$
\begin{equation*}
\Phi\left(\mathrm{c}_{\mathrm{U}}\right)-\Phi\left(\mathrm{c}_{\mathrm{L}}\right)=\Phi\left(-\frac{\delta}{\bar{\sigma}_{\mathrm{U}}}+\frac{\Delta}{\bar{\sigma}_{\mathrm{U}}}-\frac{\mathrm{Z}_{1-\alpha}}{\mathrm{w}_{\mathrm{U}}}\right)-\Phi\left(-\frac{\delta}{\bar{\sigma}_{\mathrm{L}}}-\frac{\Delta}{\bar{\sigma}_{\mathrm{L}}}+\frac{\mathrm{Z}_{1-\alpha}}{\mathrm{w}_{\mathrm{L}}}\right) \tag{Eq.53}
\end{equation*}
$$

for $\mathrm{c}_{\mathrm{U}}-\mathrm{c}_{\mathrm{L}}>0$ (0, otherwise), and

$$
\begin{aligned}
& \mathrm{a}_{\mathrm{U}}=-\delta(1+\Delta)-2\left(\mathrm{p}_{01}-\Delta\right), \\
& \mathrm{b}_{\mathrm{U}}=-\mathrm{p}_{01} \Delta(1-\Delta), \\
& \overline{\mathrm{p}}_{\mathrm{U}, 01}=\left[-\mathrm{a}_{\mathrm{U}}+\left(\mathrm{a}_{\mathrm{U}}^{2}-8 \mathrm{~b}_{\mathrm{U}}\right)^{1 / 2}\right] / 4, \\
& \bar{\sigma}_{\mathrm{U}}=\left(\frac{2 \overline{\mathrm{p}}_{\mathrm{U}, 01}+\Delta-\Delta^{2}}{\mathrm{n}}\right)^{1 / 2}, \\
& \mathrm{a}_{\mathrm{L}}=-\delta(1-\Delta)-2\left(\mathrm{p}_{01}+\delta\right),
\end{aligned}
$$

$$
\begin{aligned}
& \mathrm{b}_{\mathrm{L}}=\mathrm{p}_{01} \Delta(1+\Delta) \\
& \overline{\mathrm{p}}_{\mathrm{L}, 01}=\left[-\mathrm{a}_{\mathrm{L}}+\left(\mathrm{a}_{\mathrm{L}}^{2}-8 \mathrm{~b}_{\mathrm{L}}\right)^{1 / 2}\right] / 4, \\
& \mathrm{w}_{\mathrm{U}}=\mathrm{w}_{\mathrm{L}}=1 \\
& \bar{\sigma}_{\mathrm{L}}=\left(\frac{2 \overline{\mathrm{p}}_{\mathrm{L}, 01}-\Delta-\Delta^{2}}{\mathrm{n}}\right)^{1 / 2} .
\end{aligned}
$$

Using the property $\Phi^{-1} \Phi(\mathrm{x})=\mathrm{x}$, the observed significance level, algebraically solved from Equation 53, equals $\alpha=a_{L}+a_{U}$ and is computed with the functions

$$
\begin{align*}
& \alpha_{\mathrm{L}}=1-\Phi\left[\mathrm{w}_{\mathrm{L}}\left\{\Phi^{-1}\left(\Phi\left[\mathrm{c}_{\mathrm{U}}\right]-(1-\beta)\right)+\frac{\delta}{\bar{\sigma}_{\mathrm{L}}}+\frac{\Delta}{\bar{\sigma}_{\mathrm{L}}}\right\}\right]  \tag{Eq.54}\\
& \alpha_{\mathrm{U}}=1-\Phi\left[\mathrm{w}_{\mathrm{U}}\left\{-\Phi^{-1}\left(\Phi\left[\mathrm{c}_{\mathrm{L}}\right]+(1-\beta)\right)-\frac{\delta}{\bar{\sigma}_{\mathrm{U}}}+\frac{\Delta}{\bar{\sigma}_{\mathrm{U}}}\right\}\right] \tag{Eq.55}
\end{align*}
$$

Equivalence Method 2. Liu, Hsueh, Hsieh, and Chen (2002) derived a Restricted Maximum Likelihood (RMLE) power function based on the difference in discordant probabilities for the following equivalence hypothesis set for paired binomial events.

$$
\begin{aligned}
& \mathrm{H}_{0}: \mathrm{p}_{10}-\mathrm{p}_{01} \geq \Delta \text { or } \mathrm{p}_{10}-\mathrm{p}_{01} \leq-\Delta \\
& \mathrm{H}_{1}:-\Delta<\mathrm{p}_{10}-\mathrm{p}_{01}<\Delta
\end{aligned}
$$

The power function is

$$
\begin{equation*}
\Phi\left(\mathrm{c}_{\mathrm{U}}\right)-\Phi\left(\mathrm{c}_{\mathrm{L}}\right)=\Phi\left(-\frac{\delta}{\bar{\sigma}_{\mathrm{U}}}+\frac{\Delta}{\bar{\sigma}_{\mathrm{U}}}-\frac{\mathrm{Z}_{1-\alpha}}{\mathrm{w}_{\mathrm{U}}}\right)-\Phi\left(-\frac{\delta}{\bar{\sigma}_{\mathrm{L}}}-\frac{\Delta}{\bar{\sigma}_{\mathrm{L}}}+\frac{\mathrm{Z}_{1-\alpha}}{\mathrm{w}_{\mathrm{L}}}\right) \tag{Eq.56}
\end{equation*}
$$

for $\quad c_{U}-c_{L}>0(0$, otherwise $)$, and

$$
\mathrm{a}_{\mathrm{U}}=-\delta(1+\Delta)-2\left(\mathrm{p}_{01}-\Delta\right)
$$

$$
\begin{aligned}
& \mathrm{b}_{\mathrm{U}}=-\mathrm{p}_{01} \Delta(1-\Delta), \\
& \overline{\mathrm{p}}_{\mathrm{U}, 01}=\left[-\mathrm{a}_{\mathrm{U}}+\left(\mathrm{a}_{\mathrm{U}}^{2}-8 \mathrm{~b}_{\mathrm{U}}\right)^{1 / 2}\right] / 4, \\
& \mathrm{w}_{\mathrm{U}}=\frac{\left(2 \mathrm{p}_{01}+\delta-\delta^{2}\right)^{1 / 2}}{\left(2 \overline{\mathrm{p}}_{\mathrm{U}, 01}+\Delta-\Delta^{2}\right)^{1 / 2}}, \\
& \bar{\sigma}_{\mathrm{U}}=\left(\frac{2 \overline{\mathrm{p}}_{\mathrm{U}, 01}+\Delta-\Delta^{2}}{\mathrm{n}}\right)^{1 / 2}, \\
& \mathrm{a}_{\mathrm{L}}=-\delta(1-\Delta)-2\left(\mathrm{p}_{01}+\delta\right), \\
& \mathrm{b}_{\mathrm{L}}=\mathrm{p}_{01} \Delta(1+\Delta), \\
& \overline{\mathrm{p}}_{\mathrm{L}, 01}=\left[-\mathrm{a}_{\mathrm{L}}+\left(\mathrm{a}_{\mathrm{L}}^{2}-8 \mathrm{~b}_{\mathrm{L}}\right)^{1 / 2}\right] / 4, \\
& \mathrm{w}_{\mathrm{L}}=\frac{\left(2 \mathrm{p}_{01}+\delta-\delta^{2}\right)^{1 / 2}}{\left(2 \overline{\mathrm{p}}_{\mathrm{L}, 01}-\Delta-\Delta^{2}\right)^{1 / 2}}, \\
& \bar{\sigma}_{\mathrm{L}}=\left(\frac{2 \overline{\mathrm{p}}_{\mathrm{L}, 01}-\Delta-\Delta^{2}}{\mathrm{n}}\right)^{1 / 2}
\end{aligned}
$$

The observed significance level, algebraically solved from Equation 56, equals $\alpha=\mathrm{a}_{\mathrm{L}}+\mathrm{a}_{\mathrm{U}}$ and is computed with equations 54 and 55 with the RMLE estimation used for the parameters.

## Pharmacokinetics

Inference and power analysis for this study was based on statistical methodology borrowed from pharmacokinetics. This field is principally concerned with measuring the absorption, concentration, metabolization, and excretion rates of drugs in the human body (ICH E8, 1997). The physiological behaviors of pharmaceutical compounds are typically evaluated with the area under their relationship with time $t$ (e.g., hours) for each healthy volunteer. Figure 4 provides an example of a pharmacokinetic curve from Yeh and Kwan (1978). The $y$-axis represents a person's drug concentration during the absorption phase over time $t$ ( $x$-axis) measured in hours.


Figure 4. Example Pharmacokinetic Curve

For this study, this pharmacokinetic methodology was used for statistically comparing the IRFs for the $k$-sample matched study design. As a result, the $x$-axis denoted the predisposition of patients experiencing AEs. The $y$-axis represented the probability of this predisposition.

## AUC Estimation

Approaches. Two approaches for estimating AUC are available for the $k$-sample matched study design. The Partial approach (Navarro-Fontestad et al., 2005) was used to estimate the area under IRFs for each type of AE or item. The Partial estimate equaled the average area under the IRFs. The Bailer approach (Nedelman, Gibiansky, \& Lau, 1995) takes the opposite direction for AUC estimation. Measures of patient predisposition for experiencing a particular type of AE are first averaged across types of AEs, resulting in one IRF per treatment group. The Bailer AUC estimate then equaled the area under this single IRF.

Approximations. The linear trapezoid and spline approximations to the exact area under IRFs was investigated in this study. The linear trapezoid approximation assumes that the AUC is based on the functional form $\mathrm{y}=\mathrm{a}+\mathrm{bt}$, where $a$ and $b$ are linear (Yeh, 2002). This approximation, based on two incremental points per subinterval, is calculated as follows

$$
\begin{equation*}
\hat{\mathrm{A}}=\sum_{\mathrm{k}=0}^{\mathrm{K}} \mathrm{w}_{\mathrm{k}} \overline{\mathrm{y}}_{\mathrm{k}} \tag{Eq.63}
\end{equation*}
$$

where $w_{k}=\left\{\begin{array}{l}\frac{t_{1}-0}{2}, k=0 \\ \frac{t_{k+1}-t_{k-1}}{2}, k=1, \ldots, K-1 . \\ \frac{t_{K}-t_{K-1}}{2}, k=K\end{array}\right.$
The spline approximation (Yeh \& Kwan, 1978) assumes the AUC is based on the functional form

$$
\mathrm{y}=\frac{\partial^{2}}{\partial^{2} \mathrm{t}}[\mathrm{y}]=\frac{\partial^{2}}{\partial^{2} \mathrm{t}}\left[\mathrm{a}_{\mathrm{i}}+\mathrm{b}_{\mathrm{i}} \mathrm{t}+\mathrm{c}_{\mathrm{i}} \mathrm{t}^{2}+\mathrm{d}_{\mathrm{i}} \mathrm{t}^{3}\right]
$$

where $\frac{\partial^{2}}{\partial^{2} \mathrm{t}}[\cdot]$ denotes the second derivative of the function $y$ with respect to parameter $t$.

This approximation, based on three incremental points per subinterval, is generally calculated as follows

$$
\begin{equation*}
\hat{A}=\sum_{i=1}^{n}\left\{\frac{h_{i}^{3}}{24}\left(\ddot{y}_{\mathrm{i}}+\ddot{y}_{\mathrm{i}-1}\right)+\mathrm{h}_{\mathrm{i}}\left[\frac{1}{2} \mathrm{~s}_{1}\left(\mathrm{t}_{\mathrm{i}}+\mathrm{t}_{\mathrm{i}}-1\right)+\mathrm{s}_{2}\right]\right\} \tag{Eq.64}
\end{equation*}
$$

where $s_{1}=\frac{1}{h_{i}}\left(y_{i}-y_{i-1}\right)-\frac{h_{i}}{6}\left(\ddot{y}_{i}-\ddot{y}_{i-1}\right)$ and

$$
\mathrm{s}_{2}=\frac{1}{\mathrm{~h}_{\mathrm{i}}}\left(\mathrm{t}_{\mathrm{i}} \mathrm{y}_{\mathrm{i}-1}-\mathrm{y}_{\mathrm{i}} \mathrm{t}_{\mathrm{i}-1}\right)-\frac{\mathrm{h}_{\mathrm{i}}}{6}\left(\mathrm{t}_{\mathrm{i}} \ddot{\mathrm{y}}_{\mathrm{i}}-1-\ddot{\mathrm{y}}_{\mathrm{i}-1} \mathrm{t}_{\mathrm{i}-1}\right) .
$$

The parameters $\ddot{y}_{i}$ and $\ddot{y}_{i-1}$ are obtained by solving the following system of linear equations

$$
\left\{\begin{array}{l}
\frac{h_{i-1}}{6} \ddot{y}_{i-2}+\frac{1}{3}\left(h_{i}+h_{i-1}\right) \ddot{y}_{i-1}+\frac{h_{i}}{6} \ddot{y}_{i}=\frac{y_{i}-y_{i-1}}{h_{i}}-\frac{y_{i-1}-y_{i-2}}{h_{i-1}} \\
\frac{1}{h_{i-1}} \ddot{y}_{i-2}-\left(\frac{1}{h_{i-1}}+\frac{1}{h_{i}}\right) \ddot{y}_{i-1}+\frac{1}{h_{i}} \ddot{y}_{i}=0 \\
\frac{h_{i-1}}{6} \ddot{y}_{i-2}+\frac{h_{i-1}}{3} \ddot{y}_{i-1}+0 \cdot \ddot{y}_{i}=\frac{y_{i-1}}{h_{i}} \log _{e}\left(\frac{y_{i}}{y_{i-1}}\right)-\frac{y_{i-1}-y_{i-2}}{h_{i-1}}
\end{array} .\right.
$$

This system solves for the vector of unknown parameters

$$
\left[\begin{array}{lll}
\ddot{y}_{i-2} & \ddot{y}_{i-1} & \ddot{y}_{i}
\end{array}\right]^{\mathrm{T}} .
$$

For this study, the SAS (v9.2) PROC EXPAND procedure was utilized to compute the linear trapezoid and spline approximations to the exact area under IRFs. SAS 9.1.3 was used for all other analyses.

It is anticipated that if the underlying empirical curve of the observed data points (i.e., probability of AE predisposition vs. AE Predisposition $(\boldsymbol{\theta})$ ) is relatively linear, the linear trapezoid and spline approximations will be comparable. As the curve of these observations becomes more nonlinear (e.g., cubic), the spline may more closely approximate the exact AUC. A numerical comparison of these approximations will be provided in Chapter IV.

Variance of AUC Estimators. The Partial Batch, Bootstrap, and Jackknife methods for estimating the variance of the approximation to the area under IRFs was computed for each combination of estimation approaches. The Partial Batch method (Navarro-Fontestad et al., 2005) estimates the variance of the AUC as

$$
\begin{equation*}
\mathrm{S}_{\mathrm{A}_{\mathrm{i}}}^{2}=\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{~A}_{\mathrm{i}}-\overline{\mathrm{A}}\right)^{2} \tag{Eq.65}
\end{equation*}
$$

The variance across AE types is $\mathrm{V}(\hat{\mathrm{A}})=\frac{1}{\mathrm{n}_{\mathrm{k}}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{S}_{\mathrm{A}_{\mathrm{i}}}^{2}$. The Bootstrap method (Dunning, 2007) estimates the AUC variance as

$$
\begin{equation*}
\mathrm{V}(\hat{\mathrm{~A}})=\frac{1}{\mathrm{~B}-1} \sum_{\mathrm{b}=1}^{\mathrm{B}}\left(\mathrm{~A}_{\mathrm{b}}^{*}-\mathrm{A}^{*}\right)^{2} \tag{Eq.66}
\end{equation*}
$$

where $A^{*}=\frac{1}{B} \sum_{b=1}^{B} A_{b}^{*}$
for B bootstrap samples. The Jackknife method (Bonate, 1998; Hanley \& Hajian-Tilaki, 1997) estimates AUC variance as

$$
\begin{equation*}
\overline{\mathrm{V}}=\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{~s}_{\mathrm{A}_{\mathrm{i}}}^{2} \tag{Eq.67}
\end{equation*}
$$

where $V_{i}=n \hat{A}-(n-1) A_{i}$ and

$$
\mathrm{S}_{\mathrm{A}}^{2}=\frac{1}{\mathrm{n}(\mathrm{n}-1)} \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{~V}_{\mathrm{i}}-\overline{\mathrm{V}}\right)^{2}
$$

## Risk Prediction Algorithms

Risk Prediction Algorithms (RPAs) are utilized for identifying those individuals that are "at-risk" (Newman \& Roth, 2005, p. 1729) for experiencing a particular type of AE from the treatment of their medical condition(s). These algorithms can be used to investigate any type of AE, especially those that have serious health consequences. These algorithms identify levels of risk based on combinations of clinically relevant predictors of AEs (Cooper, Miller, \& Humphries, 2005). Such risk can be measured by factor scores (Bridgewater et al., 1998), probabilities, or predicted rates of AEs (Geraci et al., 1993).

The quantified risk can then be categorized into deciles (Geraci et al., 1993) or ordinal levels (e.g., low-risk, moderate-risk, and severe-risk) (Heddle, Klama, Griffith, Shukla, \& Kelton, 1993). After the development of this quantified risk, the discrimination and calibration of the RPAs may be fine-tuned and updated as necessary.

For AEs, these algorithms may be constructed in the following manner. The first step consists of using a logistic regression model to identify clinically relevant factors that are statistically predictive of the AE under investigation. The researcher should be cognizant of sample size requirements so that confounding results are avoided. The second step comprises the development of risk factor scores (Heddle et al., 1993). Numerous approaches are available for generating risk factor scores, but are generally a function on the "degree of significance" (Higgins et al., 1992, p. 2347) of indicator variables (e.g., $\mathrm{AGE}=1$ for patient ages 20-25 and AGE $=0$ otherwise). Newman and Roth (2005) used regression coefficients $(R)$ from a logistic regression model to determine the expected probability of an $A E$, where this probability equals $\mathrm{e}^{\mathrm{R}} /\left(1+\mathrm{e}^{\mathrm{R}}\right)$.

The development of risk gradation into deciles, ordinal categories, or other mechanisms occurs for the next step. Step 4 involves evaluating the calibration and discrimination of the algorithm. Wasson, Sox, Neff, and Goldman (1985) recommend testing prediction rules with bootstrap, jackknife, and split sample methodology and repeat this analysis for different populations.

Calibration for RPAs is concerned with the algorithm's goodness-of-fit. The Homer-Lemeshow statistic has been used to gauge the relation between observed and expected rates of an AE. Nonsignificant $p$-values are indicative of good model fit (Cooper et al., 2005). Discrimination is the predictive power of the algorithm. As this power
increases, the algorithm is better able to differentially predict AEs for at-risk and not-atrisk populations (Steyerberg et al., 2010). Discrimination may be evaluated with the $c$-statistic or the area under an Receiver Operator Characteristic (ROC) curve (Geraci et al., 1993). For binomial variables, the interpretation of the $c$-statistic which is the same as the area under the ROC (Bridgewater et al., 1998) is provided in Table 2.

## Table 2

## Predictive Power of Algorithms

| C-statistic | Predictive Power |
| :---: | :---: |
| $0 \leq$ value $\leq 0.5$ | None |
| $0.5<$ value $<0.7$ | Low |
| $0.7 \leq$ value $<0.9$ | Moderate |
| value $\geq 0.9$ | High |

## Application of IRT to Adverse Events

A single attempt in modeling AEs with IRT methodology is available in the literature. Zwinderman (2001) used the Twisting IRT method to generate and visually compare smoothed survival curves between neurologically-associated AEs for patients undergoing surgery for a single brain metastasis. This hybrid method is a cross-product of the polytomous IRT model

$$
\operatorname{Pr}\left(Y_{u}=j \mid \theta_{u}\right)=\frac{e^{\psi_{j}+j \theta_{i t}}}{1+\sum_{m=1}^{J} e^{\psi_{m}+m \theta_{i t}}}
$$

for patient $i$, time $t$, category $j(\mathrm{j}=1, \ldots, \mathrm{~J})$, latent trait score $\boldsymbol{\theta}$, and category-scoring parameter $\psi$, and the survivor curve

$$
\mathrm{S}\left(\mathrm{Y}_{\mathrm{i}}\right)=\exp \left\{-\sum_{\mathrm{u}=1}^{\mathrm{i}} \mathrm{~h}(\mathrm{u})\right\} .
$$

The latent trait for these two models was assumed to be neurological complaints and functional dependence on time $t$, respectively. The cumulative hazards function, $\mathrm{h}(\mathrm{u})$, denotes the probability of a patient's time to first experiencing an AE.

Although this appears to be the first attempt to model AEs with IRT methodology, this work is not readily applicable to ICTs that are not longitudinal in data structure. This study expanded upon the ideas of Zwinderman, and developed statistical methodology for performing statistical inference on IRFs from IRT models, and power analysis for determining the number of patients required for such comparisons.

## CHAPTER III

## METHODOLOGY

## Data Sources for Study

This study focused on types of AEs that may result from the transfusion of blood products. The Centers for Disease Control (CDC) has recently developed an electronic Hemovigilance system (Appendix L) for hospitals in the United States to centrally report these events. The purpose of this electronic database, in part, is to understand the incidence of rare reactions to transfusions of whole blood or any of its components (e.g., platelets, red blood cells, etc.). The historical occurrence of these types of AEs is provided in Table 3 (p. 81). Arbitrary item locations were based on the descending order of historical incidence.

A total of $30,044,000$ blood product components were transfused in the United States in 2006. Of these components, 10,388,000 platelet products and 1,480,270 (apheresis) red blood cell products were transfused (U.S. Department of Health and Human Services, 2007b). The type of AE with the highest and lowest annual incidence for the United States was the Delayed Serologic Transfusion Reaction (DSTR) and the Hypotensive Transfusion Reaction (HTR), respectively. DSTR was projected to occur for 1 in every 152 transfusions (Ness, Shirey, Thoman, \& Buck, 1990) for an annual total of 198,291 AEs. HTR was projected to occur for 1 in every 33,333 transfusions (Khalid, Usman, \& Khurshid, 2010) for an annual total of 45 AEs.

## Design of Study

In this study, Monte Carlo (MC) simulations were utilized to construct the data for analysis. Simulations were based on the historical occurrence of AEs reported in Table 3. This section details the algorithm for constructing these datasets for the $k$-sample matched study design. The dataset for each type of AE comprised $N=500,000$ patients for paired treatments A and B. Smaller sample sizes resulted in an algorithm with insufficient quality, which will be investigated in Chapter IV. Simulated bivariate binomial AE data was evaluated in terms of how far observed AE rates deviated from the target rates, the correlation between treatments A and B, and the stability of the AE Predisposition ( $\boldsymbol{\theta})$, discrimination, and difficulty parameters from the 2-PL IRT model. Investigation of these scenarios was not performed for the remaining types of data and IRT models because comparable results and conclusions were anticipated.

## Simulation of Bivariate Binomial Data

A bivariate binomial algorithm (Van Ness, Holford, \& Dubin, 2005) based on the $2 \times 2$ contingency table data structure (Table 1) was used to simulate MC data for each type of AE (Table 3, p. 81). This approach simulated the counts of concordant pairs (c) or main diagonals $(0,0)$ and $(1,1)$ for the historical incidence rate reported. The patterns $(0,0)$ and $(1,1)$ denoted that patients in both treatment groups did not and did, respectively, experience a particular type of AE. The remaining $n-c$ counts were then randomly allocated to the discordant pairs (d) or off diagonals $(0,1)$ and $(1,0)$. With the counts of all four cells populated, analysis could then be performed.

Table 3
Historical Incidence of Transfusion-Related Adverse Events in United States

| Item / Type of Adverse Event | Blood Product | Historical Occurrence | Incidence Ratio | Annual \# of Transfusions | \# of AEs |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 / Delayed Serologic Transfusion Reaction | All blood components | $0.66 \%{ }^{1}$ | 1:152 | 30,044,000 | 198,291 |
| 2 / Allergic Reaction (including anaphylaxis) | Platelets | $0.266 \%^{2}$ | 1:376 | 10,388,000 | 27,633 |
| 3 / Delayed Hemolytic Transfusion Reaction | All blood components | $0.12 \%{ }^{1}$ | 1:833 | 30,044,000 | 36,053 |
| 4 / Infection (Bacterial Contamination) | Platelets | $0.053 \%{ }^{3}$ | 1:1887 | 10,388,000 | 5,506 |
| 5 / Febrile Non-hemolytic Transfusion Reaction | Platelets | $0.0489 \%{ }^{2}$ | 1:2045 | 10,388,000 | 5,080 |
| 6 / Transfusion Associated Circulatory Overload | Platelets | $0.0168 \%{ }^{4}$ | 1:5952 | 10,388,000 | 1,746 |
| 7 / Transfusion Associated Graft vs Host Disease | All blood components | $0.00565 \%{ }^{5}$ | 1:17699 | 30,044,000 | 1,698 |
| 8 / Acute Hemolytic Transfusion Reaction | Red blood cells | $0.004 \%{ }^{6}$ | 1:25000 | 1,480,270 | 60 |
| 9 / Hypotensive Transfusion Reaction | Red blood cells | $0.003 \%{ }^{6}$ | 1:33333 | 1,480,270 | 45 |

Note. ${ }^{1}$ Ness, Shirey, Thoman, \& Buck, 1990; ${ }^{2}$ Huh \& Lichtiger, 1995; ${ }^{3}$ Nova Scotia Department of Health, 2005; ${ }^{4}$ Kleinman, Chan, \& Robillard, 2003; ${ }^{5}$ Rühl, Bein, \& Sachs, 2009; ${ }^{6}$ Khalid, Usman, \& Khurshid, 2010.

The steps for executing this bivariate binomial algorithm are provided. The first step was to generate cell counts $x_{11 a}$ (Treatment A) and $x_{11 b}$ (Treatment B), the frequencies of patients who experienced a particular type of AE. The SAS function RANPOI used the historical incidence of each type of AE to simulate the Poisson counts $x_{11 a}$ and $x_{11 b}$. The frequency $x_{11}$ was then set to the minimum of $x_{11 a}$ and $x_{1 l b}$. The second step distributed the difference between $x_{11 a}$ and $x_{11 b}$ over the off-diagonal or discordant cells $x_{10}$ and $x_{01}$. The count $x_{10}$ denoted the frequency of patients who experienced an AE
on Treatment A but not Treatment B. The count $x_{01}$ denoted the frequency of patients who experienced an AE on Treatment B but not Treatment A. If the difference between $x_{11 a}$ and $x_{11 b}$ was even, then $x_{10}$ was set to $x_{01}$. If the difference between $x_{11 a}$ and $x_{11 b}$ was odd, then $x_{10}$ and $x_{01}$ equaled, respectively, the ceiling and floor of the average of the difference $\left(\mathrm{x}_{11 \mathrm{a}}+\mathrm{x}_{11 \mathrm{~b}}\right) / 2$. For the next step, the count $x_{00}$ equaled the total sample size (i.e., 500,000 ) minus the sum of $x_{11}, x_{10}$, and $x_{01}$. The sample size of 500,000 patients was the minimum required for simulated rates of AEs to be comparable to the historical AE rate (i.e., target rate).

An example is provided to illustrate this portion of the algorithm. Step 1 resulted in the cell counts $x_{11 a}=50$ for Treatment A and $x_{11 b}=55$ for Treatment B. The difference between $x_{11 a}$ and $x_{11 b}$ is 5 . Next, the remainder of the difference between $x_{11 a}$ and $x_{11 b}$, the number of discordant pairs, is then distributed between $x_{10}$ and $x_{01}$. Given that this difference is odd, counts of 3 and 2 are allocated to $x_{10}$ and $x_{01}$, respectively.

The SAS function RANTBL was then used to simulate $2 \times 2$ contingency tables based on the cell counts resulting from the previous four steps. The proportion of AE types for cell frequencies $x_{10}$ and $x_{11}$ was first simulated for row 1 with the function RANTBL(seed1, $\left.\mathrm{p}_{10}, \mathrm{p}_{11}\right)$. The proportion of AE types for cell frequencies $x_{00}$ and $x_{01}$ was then simulated for row 2 with the function RANTBL(seed2, $\mathrm{p}_{00}, \mathrm{p}_{01}$ ). These two rows were then stacked to construct each simulated $2 \times 2$ contingency table. Seed 1 was randomly generated for the simulations and seed 2 differed from this seed by a factor of 50. Seeds with small differences (e.g., less than 100) are known to cause dependence and serial correlation of data vectors (Bremer, Perez, Smith, \& Westfall, 2004), the data
structure that was of primary interest in this study. These assumptions were fully evaluated later in this chapter.

## Adverse Event Data Patterns

In addition to simulating AE types based on historical incidence, four patterns of characteristics (Table 4, p. 84) were investigated. These analyses reflected the paired treatment comparison of individual AE types for 250 patients for various arbitary item locations derived from combinations of the AE characteristics severity, relationship to treatment, and seriousness (Table 5, p. 84). Severity was categorized as "Non-severe" and "Severe." Relationship to treatment was categorized as "Ruled Out," "Doubtful," "Possible," "Probable," and "Definite." Seriousness was categorized as "No" and "Yes," where serious AEs were identified by "Life-threatening" and "Death" severity categories per the Case Report Form in Appendix L. Arbitrary item locations were based on the medical implications of the characteristics associated with the AE.

As the medical implication of a reaction increased for a patient, the AE location was said to increase in magnitude. AE types that were not serious, not related to treatment, and non-severe were allocated the lowest item location. This category may mimic the easiest item on a test instrument in traditional IRT application. Serious AEs that were related to treatment and were severe were assigned the largest item location. These characteristics represented the most difficult or "hardest" item (Hays et al., 2000).

Data Pattern 1 assumed that the AE characteristics were the same for both treatment groups and had occurrence of $20 \%$ and $10 \%$ for Treatment A Location 1 and Treatment B Location 1, respectively. The AE characteristics were slightly different for

Data Pattern 2, where AE occurrence was $10 \%$ for Treatment A Location 3 and 20\% for Treatment B Location 6. Data Pattern 3 assumed an extreme difference in AE characteristics, where AE occurrence was $20 \%$ for Treatment A Location 20 and 5\% for Treatment B Location 1. All other AE locations had an occurrence of $0 \%$. For the last data pattern, the characteristics for the 20 AE locations for both Treatment A and B were randomly allocated using the SAS RANBIN function with approximate occurrence of $15 \%$ to $20 \%$.

## Table 4

Patterns of Simulated Adverse Event Characteristics

| Data <br> Pattern | Treatment A |  |  | Treatment B |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Severity | Relationship | Serious | Severity | Relationship | Serious |
| 1 | Non-severe | Ruled Out | No | Non-severe | Ruled Out | No |
| 2 | Severe | Possible | No | Non-severe | Doubtful | No |
| 3 | Severe | Definite | Yes | Non-severe | Ruled Out | No |
| 4 | Random | Random | Random | Random | Random | Random |

Table 5
Item Locations for Adverse Event Characteristics

| Relation to Treatment | Severity | Location |  | Relation to Treatment | Severity | Location |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\begin{aligned} & \text { Serious } \\ & =\mathbf{N o} \end{aligned}$ | $\begin{aligned} & \text { Serious } \\ & =\text { Yes } \end{aligned}$ |  |  | $\begin{aligned} & \text { Serious } \\ & =\text { No } \end{aligned}$ | $\begin{aligned} & \text { Serious } \\ & =\text { Yes } \end{aligned}$ |
| Ruled Out | Non-severe | 1 | 11 | Possible | Severe | 6 | 16 |
| Ruled Out | Severe | 2 | 12 | Probable | Non-severe | 7 | 17 |
| Doubtful | Non-severe | 3 | 13 | Probable | Severe | 8 | 18 |
| Doubtful | Severe | 4 | 14 | Definite | Non-severe | 9 | 19 |
| Possible | Non-severe | 5 | 15 | Definite | Severe | 10 | 20 |

## Research Questions

The research questions for this study were evaluated with $1,000 \mathrm{MC}$ data sets each comprising $k=9$ types of transfusion-related AEs (Table 3, p. 81) for two dependent treatment groups. Each IRT analysis was based on simple random samples (SRS) of sizes $n=30$ and 250 patients from a total of $N=500,000$ patients (Treatment A - Treatment B). These sample sizes reflect the typical range of patients enrolled in transfusion medicine industry clinical trials (ICTs).

The first research question pertained to modeling AEs with Item Response Theory (IRT). A single attempt has been made to model AEs with IRT (Zwinderman, 2001), and no publications are available on modeling rare binomial events with this methodology. This study sought to determine whether or not IRT can be used to model rare binary events characterized as AEs. The purpose of the second research question was to fill a gap pertaining the statistical inference and power analysis of IRT models. This study introduced the concept of clinical equivalence into IRT, and the superiority objective was advanced. This study determined which forms of statistical inference and power analysis were optimal for comparing IRFs between two treatment groups where these groups were dependent. The last research question compared sample size requirements between the IRT power functions derived by the author and existing approaches for superiority and equivalence study objectives.

## Research Question 1

The first research question pertaining to the feasibility of IRT is repeated as

Can one or more forms of a two-parameter Bayesian logistic IRT model
be feasibly used to model AEs?
This study analyzed MC AE data with three Bayesian forms of the two-parameter logistic IRT model. Feasibility was evaluated with three approaches each to assess Bayesian Gibbs sampling algorithms and IRT model goodness-of-fit. If these algorithms demonstrated convergence and goodness-of-fit was reasonably achieved, IRT would be considered a viable solution for modeling rare binary events. Multiple chains of starting values (Table 6) were utilized to ensure that any conclusions reached based on the IRT parameter estimates were not an artifact of the data. These assumptions were evaluated for overall historical AE rates. Findings based on this analysis were recycled for data patterns 1-4.

Table 6
Starting Values for Gibbs Samplers

| $\boldsymbol{\theta}$ | $\mathbf{A}$ | $\mathbf{B}$ |
| :---: | :---: | :---: |
| $0, \pm 1$ | $\mu_{\mathrm{a}}=2, \mu_{\mathrm{a}} \pm 5$ | $-\left(1+\mu_{\mathrm{a}}\right)^{1 / 2} \Phi^{-1}\left(\hat{p}_{\mathrm{j}}\right)$ |

Gibbs Sampler Assumption 1. The first assumption required that the mean of each latent trait parameter "converge to the chain's stationary distribution" (Patz \& Junker, 1999b, p. 150). Convergence occurred when the mean of a parameter estimate was stable over $i$ iterations of the Gibbs sampler. Trace plots (Johnson \& Albert, 1999) and the Gelman-Rubin convergence statistic (Johnson \& Sinharay, 2005) were used to identify
the approximate location at which each parameter reached a stationary state. Trace plots are iterative and the Gelman-Rubin statistic is cumulative over the $x$-axis.

Figure 5 presents an example trace plot. In this example, the parameter estimate starts to converge at approximately 250 iterations. The period prior to 250 is known as the burn-in period, and the stationary state consists of iterations 250-1000. Statistical inference should only be based on an estimate from this stable state.


Figure 5. Example Trace Plot for 1000 Gibbs Sampler Iterations

The Gelman-Rubin (G-R) statistic uses the ratio of the total variance and within variance to identify parameter stability. This statistic equals

$$
\begin{equation*}
\hat{\mathrm{R}}^{1 / 2}=\left(\frac{\text { Total }}{\mathrm{WV}}\right)^{1 / 2} \tag{Eq.68}
\end{equation*}
$$

The within-variance (WV) is estimated as

$$
\frac{1}{\mathrm{~m}(\mathrm{n}-1)} \sum_{\mathrm{i}=1}^{\mathrm{m}} \sum_{\mathrm{j}=1}^{\mathrm{n}}\left(\lambda_{\mathrm{ij}}-\bar{\lambda}_{\mathrm{i}}\right)^{2}
$$

where $\lambda$ represents the parameter estimate, $m$ denotes the number of blocks, and $n$ denotes the length of each block. The between-variance (BV) equals

$$
\frac{\mathrm{n}}{\mathrm{~m}-1} \sum_{\mathrm{i}=1}^{\mathrm{m}}\left(\bar{\lambda}_{\mathrm{i}}-\bar{\lambda}\right)^{2}
$$

The total variance then equals

$$
\left(\frac{\mathrm{n}-1}{\mathrm{n}}\right) \mathrm{WV}+\frac{1}{\mathrm{n}}(\mathrm{BV})
$$

Stability of parameter estimates is said to occur when $\hat{\mathrm{R}}^{1 / 2}$ converges to one (Gelman, 1996).

Gibbs Sampler Assumption 2. The second assumption required that the standard deviation (SD) of each latent trait parameter represent a well-mixed density. Mixing was said to occur when the $S D$ of a parameter estimate was stable over $i$ iterations of the Gibbs sampler. Trace plots and the G-R convergence statistic were used to identify when the density is thoroughly mixed.

Gibbs Sampler Assumption 3. The last assumption, referred to as "thinning the chain" (Lynch, 2007, p. 147), requires that the estimates within a stationary state are not serially dependent. If the parameter estimates are autocorrelated (May, 2006), the resulting estimators (e.g., mean and variance) are not reliable for statistical inference. To remove the serial autocorrelation from a sequence of values, the lag resulting in nonsignificant autocorrelation needs to be identified. As an example, if nonsignificant autocorrelation occurs for lag 25 , the mean and $S D$ of the parameter should be based on a
systematic random sample (Lohr, 1999) of every $25^{\text {th }}$ value in the established stationary state. For each iterative lag, the autocorrelation function was estimated by Equation 69.

$$
\begin{equation*}
A C F_{L}=\left(\frac{T}{T-L}\right) \frac{\sum_{i=1}^{T-L}\left(x_{i}-\bar{x}\right)\left(x_{i+L}-\bar{x}\right)}{\sum_{i=1}^{T}\left(x_{i}-\bar{x}\right)^{2}} \tag{Eq.69}
\end{equation*}
$$

$x_{i}$ is the sampled value of $x$ for iteration $i$. The estimator $\overline{\mathrm{x}}$ denotes the mean of the values $x_{i}$ in interval $T$ which equals $2 l a g$. For example, if a lag is 10 , then T will contain 20 values. A nonparametric bootstrap method was utilized to construct a CI on the mean of the ACF statistic. The first lag having the CI on the mean of the ACF contain zero was identified as the lag to remove autocorrelation from the stationary state.

After stationary parameter estimates were identified from the Gibbs sampling algorithm, graphical displays and formal methods based on hypothesis testing were utilized to gauge the fit of each IRT model in terms of AE items and patients. If the Gibbs samplers and fit of these models were visually and statistically adequate, it was concluded that IRT could reasonably model rare binary event data.

Toribio (2006) argues that existing methods for determining the goodness-of-fit (GoF) of logistic IRT models overall as well as the individual fit of items and persons can be problematic. The reason for this is that "their exact distributions are unknown and the asymptotic chi-square distribution may not always be true" (p. 39). This may mean that the inference from the GoF statistics is not always reliable. As a result, this study presented visual methods for assessing fit as well as attempt at a solution to the distribution problem noted by this author.

Nonparametric bootstrap methodology (Stute, Manteiga, \& Quindmil, 1993) was utilized to assess the fit of each AE item. This approach was distribution-free and did not require the exact form the test statistic under investigation. The GoF statistic was computed for each bootstrap and compared to the observed test statistic for $n=30$ and 250. The $p$-value was based on the bootstrapped limits of the CI. Poor model fit was said to occur for $p$-values less than 0.10 (Toribio, 2006).

Posterior Probability Interval Plot. This type of plot was generated with the following sequence of steps (Sinharay, Johnson, \& Stern, 2006; Johnson \& Albert, 1999) for each type of AE:

1. Group AE Predisposition $(\boldsymbol{\theta})$ ( $x$-axis) into $S$ intervals. The number of intervals was a function of the statistic used to assess IRT model fit.
2. Calculate the mean or central AE Predisposition $(\boldsymbol{\theta})$ score of each interval $S$. This parameter was denoted as $\theta_{\mathrm{s}}^{(\mathrm{c})}$.
3. Estimate the proportion of patients within each interval $S$ that experienced the associated AE. This parameter was denoted as $\hat{\mathrm{p}}_{\mathrm{js}}$.
4. Estimate the probability of a patient experiencing an AE for each interval $S$. This parameter was denoted as $\hat{p}_{j s}^{(c)}$ and equaled $F\left(a_{j} \theta_{j s}^{(c)}-b_{j}\right)$, where $F$ denoted the two-parameter logistic IRT joint density.
5. Estimate the posterior probability interval (PPI) for each value of $\theta_{\mathrm{s}}^{(\mathrm{c})}$ per interval $S$. This interval equaled the mean of the AE Predisposition $(\boldsymbol{\theta})$ scores
within interval $S$ plus and minus the normal critical value times the standard error $(S E)$ of the AE Predisposition $(\boldsymbol{\theta})$ scores within interval $S$.

$$
\begin{equation*}
\bar{\theta}_{\mathrm{s}} \pm \mathrm{Z}_{1-\alpha / 2} \operatorname{SE}\left(\theta_{\mathrm{s}}\right) \tag{Eq.70}
\end{equation*}
$$

6. Plot the IRF through the mean AE Predisposition $(\boldsymbol{\theta})$ score of each interval $S$.
7. Next, overlay the PPI onto this fitted curve.
8. Finally, overlay the observed AE Predisposition ( $\boldsymbol{\theta})$ scores onto this graph. IRT model fit was then visually assessed by examining the location of the observed AE Predisposition ( $\boldsymbol{\theta}$ ) scores in relation to the PPI. Reasonable model fit was evidenced by the observed AE Predisposition $(\boldsymbol{\theta})$ scores being located within the boundaries of the PPI.

Using the above steps, Bayesian residuals were computed as

$$
\begin{equation*}
\mathrm{r}_{\mathrm{ij}}=\hat{\mathrm{p}}_{\mathrm{js}}-\mathrm{p}_{\mathrm{js}}^{(\mathrm{c})} \tag{Eq.71}
\end{equation*}
$$

The nonparametric bootstrap Wald CI ( $t$-distribution) was computed for each patient. CIs close to zero represented reasonable model fit.

Although these graphs were utilized to gauge IRT model fit, some common inferential approaches were discussed below. Bock's Index was used to assess model fit for each AE type. This GoF statistic has the form

$$
\begin{equation*}
B_{k j}=\sum_{\mathrm{s}=1}^{\mathrm{S}} \frac{\left(\mathrm{f}_{\mathrm{js}}-\mathrm{f}_{\mathrm{s}} \hat{\mathrm{p}}_{\mathrm{js}}\right)^{2}}{\mathrm{f}_{\mathrm{s}} \hat{\mathrm{p}}_{\mathrm{js}} \hat{\mathrm{q}}_{\mathrm{js}}} \tag{Eq.72}
\end{equation*}
$$

In this study, the parameter $f_{s}$ was the number of patients in subinterval $s$. The parameter $f_{j s}$ was the number of patients who experienced AE type $j$ in subinterval $s$. The
parameter $\hat{\mathrm{p}}_{\mathrm{js}}$ equaled $\mathrm{e}^{-v} /\left(1+\mathrm{e}^{-v}\right)$, where $v=\mathrm{a}_{\mathrm{j}}\left(\theta_{\mathrm{s}}-\mathrm{b}_{\mathrm{j}}\right)$ for $\boldsymbol{\theta}_{s}$, the median of AE Predisposition $(\boldsymbol{\theta})$ scores for subinterval $s$, and $\hat{\mathrm{q}}_{\mathrm{js}}=1-\hat{\mathrm{p}}_{\mathrm{js}}$.

For the Bock's Index (BI) statistic, patients were grouped into subintervals $s$ for AE Predisposition $(\boldsymbol{\theta})$. Limited guidance appears to be available for how many patients should constitute each subinterval $s$, but Yen (1981) recommended a total of 10 subintervals. Furthermore, Bock (1972) recommended that the BI statistic be tested against $S-m$ degrees of freedom for the chi-square distribution, where $m$ is the number of estimated parameters. In this study, these parameters were the latent trait $\boldsymbol{Z}, \mathrm{AE}$ Predisposition $(\boldsymbol{\theta})$, discrimination $\boldsymbol{A}$, and difficulty $\boldsymbol{B}$.

Last, the W-statistic (Wright \& Stone, 1979) has been commonly used to assess the fit of each respondent. This statistic has the form

$$
\begin{equation*}
\mathrm{W}_{\mathrm{i}}=\frac{\sum_{\mathrm{j}=1}^{\mathrm{k}}\left(\mathrm{Y}_{\mathrm{ij}}-\mathrm{p}_{\mathrm{ij}}\right)^{2}}{\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{p}_{\mathrm{ij}} \mathrm{q}_{\mathrm{ij}}} \tag{Eq.73}
\end{equation*}
$$

In terms of this study, $Y_{i j}$ would represent the AE status $(0=\mathrm{AE}$ did not occur, $1=\mathrm{AE}$ occurred) for patient $i$ and type of AE $j$. The parameter $p_{i j}$ would equal $\Phi\left(\mathrm{a}_{\mathrm{j}} \theta_{\mathrm{i}}-\mathrm{b}_{\mathrm{j}}\right)$, where $a_{j}$ and $b_{j}$ were the discrimination and difficulty values for item $k$ and $\theta_{i}$ was AE Predisposition $(\boldsymbol{\theta})$ for patient $i$.

For an ICT, computation of this statistic would have limited value. The FDA does not allow the omission of any AE data from the study analysis for any reason. As a result, this statistic was not investigated in this study.

Toribio (2006) cautions that although the above GoF inferential methods are commonly applied to Bayesian estimation, they should not be utilized because these methods require the ability parameter $\boldsymbol{\theta}$ to be known. In Bayesian IRT analysis, this parameter is unknown until estimation. In addition, the distribution of the test statistics may not be chi-square, resulting in inaccurate conclusions. Furthermore, it is clear from the formula that if many patients do not experience an AE the numerator will become very large. As a result, the BI statistic may not be optimal for the data being investigated for this study. Given this information, the PPI and residual plots served as the primary source for evaluating IRT model fit.

## Research Question 2

The second research question is repeated as
For superiority and equivalence study objectives, which combination of statistics resulted in the best statistical inference as defined by

- minimum standard error of the effect?
- smallest bias in AUC approximations?
- confidence interval on the effect that achieves the highest coverage?

This research question was concerned with the statistical inference of parameters associated with the latent trait Transfusion-Related Adverse Event. Hypotheses were constructed to compare paired IRFs estimated with the 2-PL, 2-PL EX, and 2-PL MEX IRT models. Statistical inference was based on superiority and equivalence study objectives.

The Area-Under-the-Curve (AUC) statistic was utilized because it is readily applicable to empirical distributions with known and unknown form (Dunning, 2007). The linear trapezoid and spline approximations to the exact area under paired IRFs were investigated. For each combination of the AUC approximation, estimator, and variance estimate, the quality of inference (i.e., bias, mean, and variance of effects, test statistics, CI with Type I error assessment) was evaluated on simulated MC data. This study recommended the combination of estimation approaches that resulted in the best statistical inference.

Statistical Inference. The statistical information needed to compare IRFs for the $k$ sample matched study design is presented in this section. For the superiority and equivalence study objectives, the author derived the parametric and nonparametric effect, and the variability, test statistic, and CI of these effects.

The Mean Square Error (MSE) should be substituted for individual variances when statistics such as AUC are biased. As the number of patients ( $n$ ) converges to $\infty$, the sample AUC will theoretically converge to the true AUC. For small sample sizes, AUC statistics tend to be biased, meaning that the difference between the exact AUC and its approximation is not zero. The size of this bias further depends upon the functional form of the AUC approximation and sample size or the number of patients with latent trait ability scores. The bias in the AUC estimates will be investigated in Chapter IV.

Superiority. The expected value of the paired $\operatorname{effect} \omega=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}$, where $T$ denotes Treatment A and $R$ denotes Treatment B, using expectation theory, was derived

$$
\begin{align*}
& \mathrm{E}[\hat{\omega}]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}\right] \\
& \mathrm{E}[\hat{\omega}]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}\right]-\mathrm{E}\left[\mathrm{~A}_{\mathrm{R}}\right] \\
& \mathrm{E}[\hat{\omega}]=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}} \tag{Eq.74}
\end{align*}
$$

For parametric analysis, the mean of this effect was computed. For nonparametric analysis, the median of this effect was computed as

$$
\hat{M}=\operatorname{median}\{\omega\}=\left\{\begin{array}{cc}
Y_{(n+1) / 2} & n \text { odd }  \tag{Eq.75}\\
Y_{n / 2}+Y_{n / 2+1} & n \text { even }
\end{array}\right.
$$

The MSE of the paired effect $\omega=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}$, using expectation theory, was derived as

$$
\begin{aligned}
& \operatorname{MSE}[\hat{\omega}]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}[\hat{\omega}]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\mathrm{V}\left[\mathrm{~A}_{\mathrm{R}}\right]-2 \operatorname{COV}\left[\mathrm{~A}_{\mathrm{T}}, \mathrm{~A}_{\mathrm{R}}\right]
\end{aligned}
$$

For biased estimators, we have

$$
\begin{aligned}
& \operatorname{MSE}[\hat{\omega}]=\left\{\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\left(\operatorname{Bias}_{\mathrm{T}}\right)^{2}\right\}+\left\{\mathrm{V}\left[\mathrm{~A}_{\mathrm{R}}\right]+\left(\operatorname{Bias}_{\mathrm{R}}\right)^{2}\right\} \\
&- 2 \rho_{\mathrm{TR}} \sqrt{\left.\left\{\mathrm{~V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\left(\operatorname{Bias}_{\mathrm{T}}\right)^{2}\right\} \mathrm{V}\left[\mathrm{~A}_{\mathrm{R}}\right]+\left(\operatorname{Bias}_{\mathrm{R}}\right)^{2}\right\}} \\
& \operatorname{MSE}[\hat{\omega}]=\operatorname{MSE}_{\mathrm{T}}+\operatorname{MSE}_{\mathrm{R}}-2 \rho_{\mathrm{TR}}\left[\mathrm{MSE}_{\mathrm{T}} \mathrm{MSE}_{\mathrm{R}}\right]^{1 / 2}
\end{aligned}
$$

The standard error ( $S E$ ) of the effect then equaled

$$
\begin{equation*}
\operatorname{SE}[\hat{\omega}]=\left(\operatorname{MSE}[\omega] / \mathrm{n}_{\omega}\right)^{1 / 2} \tag{Eq.76}
\end{equation*}
$$

For nonparametric analysis, the $S E$ (Olive, 2005) of the effect equaled

$$
\begin{equation*}
\text { SE(effect })=0.5\left(\mathrm{Y}_{\mathrm{U}_{\mathrm{n}}}-\mathrm{Y}_{\mathrm{L}_{\mathrm{n}+1}}\right) \tag{Eq.77}
\end{equation*}
$$

$\mathrm{L}_{\mathrm{n}}$ denoted the lower order statistic (OS) of a sorted data vector, and was estimated as $\mathrm{L}_{\mathrm{n}}=\left\lceil\frac{\mathrm{n}}{2}\right\rceil-\left\lceil\sqrt{\frac{\mathrm{n}}{4}}\right\rceil . \mathrm{U}_{\mathrm{n}}$ denotes the upper OS of this vector, and was estimated as $\mathrm{U}_{\mathrm{n}}=\mathrm{n}-\mathrm{L}_{\mathrm{n}}$.

Well-known statistical properties used to derive Equations 76 and 77 were (1) $\operatorname{Var}(\mathrm{aX})=\mathrm{a}^{2} \operatorname{Var}(\mathrm{X}),(2) \operatorname{COV}(\mathrm{aX}, \mathrm{bY})=\operatorname{abCOV}(\mathrm{X}, \mathrm{Y}),(3) \hat{\rho}_{\mathrm{TR}}=$ Pearson correlation coefficient between two continuous random variables, and (4) $s_{i}^{2}(i=T, R)$ was calculated as presented above for each estimator.

The test statistic on the mean of the paired effect $\omega=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}$ was derived as

$$
\begin{equation*}
\mathrm{t}_{\text {obs }} \sim \overline{\mathrm{X}}_{\text {effect }} / \mathrm{SE}_{\text {effect }} \tag{Eq.78}
\end{equation*}
$$

This test statistic is based on the inverse $t$ CDF for $100(1-\alpha) \%$ confidence with $n-1$ degrees of freedom. The test statistic on the median of the paired effect $\omega=A_{T}-A_{R}$ was derived as

$$
\begin{equation*}
\mathrm{t}_{\mathrm{obs}} \sim \frac{\mathrm{~A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}}{\mathrm{SE}(\mathrm{M})} \tag{Eq.79}
\end{equation*}
$$

This test statistic is based on the inverse $t$ CDF for $100(1-\alpha) \%$ confidence with $\mathrm{p}=\mathrm{U}_{\mathrm{n}}-\mathrm{L}_{\mathrm{n}}-1$ degrees of freedom. The $p$-value (two-sided) was then computed with the inverse $t$ CDF as $2 \mathrm{~T}^{-1}\left(\mathrm{t}_{\text {obs }}\right)$ for $\mathrm{t}_{\text {obs }}<0$ and $2\left(1-\mathrm{T}^{-1}\left(\mathrm{t}_{\mathrm{obs}}\right)\right)$ for $\mathrm{t}_{\text {obs }} \geq 0$ using the PROBT function in SAS.

The Wald $100(1-\alpha) \%$ confidence interval on the mean of the paired effect $\omega=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}$ was derived as

$$
\begin{equation*}
\left(\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}\right) \pm \mathrm{t}_{1-\alpha / 2 ; \mathrm{n}-1}\left[\frac{\mathrm{~s}_{\mathrm{T}}^{2}+\mathrm{s}_{\mathrm{R}}^{2}-2 \rho_{\mathrm{TR}} \mathrm{~s}_{\mathrm{T}} \mathrm{~s}_{\mathrm{R}}}{\mathrm{n}}\right]^{1 / 2} \tag{Eq.80}
\end{equation*}
$$

The Wald $100(1-\alpha) \%$ confidence interval on the median of the paired effect $\omega=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}$ was derived as

$$
\begin{equation*}
\left(A_{T}-A_{R}\right) \pm t_{1-\alpha / 2 ; p} \cdot S E(\text { effect }) \tag{Eq.81}
\end{equation*}
$$

For Equation 81, the degrees of freedom $p$ equals $U_{n}-L_{n}-1$, where $U_{n}$ and $L_{n}$ were defined for Equation 77 (p. 95).

For Equations 80 and 81, the coverage (i.e., observed confidence level) was computed with 10,000 nonparametric bootstrap simulations for each AE type and SRS of $\operatorname{sizes} \mathrm{n}=30$ and 250.

Equivalence. The expected value of the lower paired effect $\omega_{1}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1+\Delta)$, using expectation theory, was derived as

$$
\begin{align*}
& \mathrm{E}\left[\hat{\omega}_{1}\right]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}\right]-\mathrm{E}\left[\mathrm{~A}_{\mathrm{R}}(1+\Delta)\right] \\
& \mathrm{E}\left[\hat{\omega}_{1}\right]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}\right]-(1+\Delta) \mathrm{E}\left[\mathrm{~A}_{\mathrm{R}}\right] \\
& \mathrm{E}\left[\hat{\omega}_{1}\right]=\mathrm{A}_{\mathrm{T}}-(1+\Delta) \mathrm{A}_{\mathrm{R}} \tag{Eq.82}
\end{align*}
$$

The expected value of the upper paired effect $\omega_{2}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)$ was derived as

$$
\begin{align*}
& \mathrm{E}\left[\hat{\omega}_{2}\right]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)\right] \\
& \mathrm{E}\left[\hat{\omega}_{2}\right]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}\right]-\mathrm{E}\left[\mathrm{~A}_{\mathrm{R}}(1-\Delta)\right] \\
& \mathrm{E}\left[\hat{\omega}_{2}\right]=\mathrm{E}\left[\mathrm{~A}_{\mathrm{T}}\right]-(1-\Delta) \mathrm{E}\left[\mathrm{~A}_{\mathrm{R}}\right] \\
& \mathrm{E}\left[\hat{\omega}_{2}\right]=\hat{\mathrm{A}}_{\mathrm{T}}-(1-\Delta) \hat{\mathrm{A}}_{\mathrm{R}} \tag{Eq.83}
\end{align*}
$$

The MSE of the lower paired effect $\omega_{1}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1+\Delta)$, using expectation theory, was derived as

$$
\begin{aligned}
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}-(1+\Delta) \mathrm{A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\mathrm{V}\left[(1+\Delta) \mathrm{A}_{\mathrm{R}}\right]-2 \operatorname{COV}\left[\mathrm{~A}_{\mathrm{T}},(1+\Delta) \mathrm{A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+(1+\Delta)^{2} \mathrm{~V}\left[\mathrm{~A}_{\mathrm{R}}\right]-(1+\Delta) 2 \operatorname{COV}\left[\mathrm{~A}_{\mathrm{T}}, \mathrm{~A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\mathrm{s}_{\mathrm{T}}^{2}+(1+\Delta)^{2} \mathrm{~s}_{\mathrm{R}}^{2}-2(1+\Delta) \rho_{\mathrm{TR}} \mathrm{~s}_{\mathrm{T}} \mathrm{~s}_{\mathrm{R}}
\end{aligned}
$$

For biased estimators (e.g., AUC), we have

$$
\begin{aligned}
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\mathrm{s}_{\mathrm{T}}^{2}+(1+\Delta)^{2} \mathrm{~s}_{\mathrm{R}}^{2}-2(1+\Delta) \rho_{\mathrm{TR}} \mathrm{~s}_{\mathrm{T}} \mathrm{~s}_{\mathrm{R}} \\
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\left\{\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\left(\mathrm{Bias}_{\mathrm{T}}\right)^{2}\right\}+(1+\Delta)^{2}\left\{\mathrm{~V}\left[\mathrm{~A}_{\mathrm{R}}\right]+\left(\mathrm{Bias}_{\mathrm{R}}\right)^{2}\right\} \\
& -2(1+\Delta) \rho_{\mathrm{TR}} \sqrt{\left.\left\{\mathrm{~V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\left(\operatorname{Bias}_{\mathrm{T}}\right)^{2}\right\} \mathrm{V}\left[\mathrm{~A}_{\mathrm{R}}\right]+\left(\mathrm{Bias}_{\mathrm{R}}\right)^{2}\right\}} \\
& \operatorname{MSE}\left[\hat{\omega}_{1}\right]=\operatorname{MSE}_{\mathrm{T}}+(1+\Delta)^{2} \mathrm{MSE}_{\mathrm{R}}-2(1+\Delta) \hat{\rho}_{\mathrm{TR}} \sqrt{\mathrm{MSE}_{\mathrm{T}} \mathrm{MSE}_{\mathrm{R}}}
\end{aligned}
$$

The $S E$ of this effect then equaled

$$
\begin{equation*}
\operatorname{SE}\left[\hat{\omega}_{1}\right]=\left(\operatorname{MSE}\left[\omega_{1}\right] / \mathrm{n}_{\omega_{1}}\right)^{1 / 2} \tag{Eq.84}
\end{equation*}
$$

The MSE of the upper paired effect $\omega_{2}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)$ was derived as

$$
\begin{aligned}
& \operatorname{MSE}\left[\hat{\omega}_{2}\right]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}-(1-\Delta) \mathrm{A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}\left[\hat{\omega}_{2}\right]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+\mathrm{V}\left[(1-\Delta) \mathrm{A}_{\mathrm{R}}\right]-2 \operatorname{COV}\left[\mathrm{~A}_{\mathrm{T}},(1-\Delta) \mathrm{A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}\left[\hat{\omega}_{2}\right]=\mathrm{V}\left[\mathrm{~A}_{\mathrm{T}}\right]+(1-\Delta)^{2} \mathrm{~V}\left[\mathrm{~A}_{\mathrm{R}}\right]-(1-\Delta) 2 \operatorname{COV}\left[\mathrm{~A}_{\mathrm{T}}, \mathrm{~A}_{\mathrm{R}}\right] \\
& \operatorname{MSE}\left[\hat{\omega}_{2}\right]=\mathrm{s}_{\mathrm{T}}^{2}+(1-\Delta)^{2} \mathrm{~s}_{\mathrm{R}}^{2}-2(1-\Delta) \rho_{\mathrm{TR}} \mathrm{~s}_{\mathrm{T}} \mathrm{~s}_{\mathrm{R}}
\end{aligned}
$$

For biased estimators, we have

$$
\operatorname{MSE}\left[\hat{\omega}_{2}\right]=\operatorname{MSE}_{\mathrm{T}}+(1-\Delta)^{2} \mathrm{MSE}_{\mathrm{R}}-2(1-\Delta) \rho_{\mathrm{TR}} \sqrt{\mathrm{MSE}_{\mathrm{T}} \mathrm{MSE}_{\mathrm{R}}}
$$

The $S E$ of this effect then equals

$$
\begin{equation*}
\operatorname{SE}\left[\hat{\omega}_{2}\right]=\left(\operatorname{MSE}\left[\omega_{2}\right] / \mathrm{n}_{\omega_{2}}\right)^{1 / 2} \tag{Eq.85}
\end{equation*}
$$

The general form of the test statistic to be used in this study was

$$
\begin{equation*}
\mathrm{t} \sim \frac{\bar{X}-\mu}{\sigma / \sqrt{n}}=\frac{\mathrm{E}[\omega]-\omega}{[\operatorname{MSE}[\omega] / n]^{1 / 2}}=\frac{\mathrm{E}[\omega]-\omega}{\operatorname{SE}[\omega]} \tag{Eq.86}
\end{equation*}
$$

where $\omega=0$. The test statistic, using distribution theory, for the mean of the lower paired effect $\omega_{1}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1+\Delta)$ was derived as

$$
\begin{equation*}
\mathrm{t}_{\omega_{1}} \sim \frac{\mathrm{~A}_{\mathrm{T}}-(1+\Delta) \mathrm{A}_{\mathrm{R}}}{\left[\frac{\mathrm{MSE}_{\mathrm{T}}+(1+\Delta)^{2} \mathrm{MSE}_{\mathrm{R}}-2(1+\Delta) \mathrm{p}_{\mathrm{TR}} \sqrt{\mathrm{MSE}_{\mathrm{T}} \mathrm{MSE}_{\mathrm{R}}}}{\mathrm{n}}\right]^{1 / 2}} \tag{Eq.87}
\end{equation*}
$$

The test statistic for the mean of the upper paired effect $\omega_{2}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)$ was derived as

$$
\begin{equation*}
\mathrm{t}_{\omega_{2}} \sim \frac{\mathrm{~A}_{\mathrm{T}}-(1-\Delta) \mathrm{A}_{\mathrm{R}}}{\left[\frac{\mathrm{MSE}_{\mathrm{T}}+(1-\Delta)^{2} \mathrm{MSE}_{\mathrm{R}}-2(1-\Delta) \rho_{\mathrm{TR}} \sqrt{\mathrm{MSE}_{\mathrm{T}} \mathrm{MSE}_{\mathrm{R}}}}{\mathrm{n}}\right]^{1 / 2}} \tag{Eq.88}
\end{equation*}
$$

Equations 87 and 88 are based on the inverse $t$ CDF employed $100(1-\alpha) \%$ confidence with $n-l$ degrees of freedom. The test statistic for the median of the lower paired effect $\omega_{1}$ was derived as

$$
\begin{equation*}
\mathrm{t}_{\omega_{1}} \sim \frac{\mathrm{M}_{\mathrm{A}_{\mathrm{T}}-(1+\Delta) \mathrm{A}_{\mathrm{R}}}}{\mathrm{SE}(\mathrm{M})} \tag{Eq.89}
\end{equation*}
$$

The test statistic for the median of the upperpaired effect $\omega_{2}$ was derived as

$$
\begin{equation*}
\mathrm{t}_{\omega_{2}} \sim \frac{\mathrm{M}_{\mathrm{A}_{\mathrm{T}}-(1-\Delta) \mathrm{A}_{\mathrm{R}}}}{\mathrm{SE}(\mathrm{M})} \tag{Eq.90}
\end{equation*}
$$

Equations 89 and 90 are based on the inverse $t$ CDF employed $100(1-\alpha) \%$ confidence with $p=U_{n}-L_{n}-1$ degrees of freedom, where $U_{n}$ and $L_{n}$ were defined for Equation 77 (p. 95).

The Wald $100(1-\alpha / 2) \%$ upper confidence limit on the mean of the lower paired effect $\omega_{1}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1+\Delta)$ was derived as

$$
\begin{equation*}
\overline{\mathrm{X}}_{\omega_{1}}+\mathrm{t}_{1-\alpha / 2 ; \mathrm{n}-1}\left[\operatorname{MSE}\left(\omega_{1}\right)\right]^{1 / 2} \tag{Eq.91}
\end{equation*}
$$

The Wald $100(1-\alpha / 2) \%$ lower confidence limit on the mean of the upper paired effect $\omega_{2}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)$ was derived as

$$
\begin{equation*}
\overline{\mathrm{X}}_{\omega_{2}}-\mathrm{t}_{1-\alpha / 2 ; \mathrm{n}-1}\left[\operatorname{MSE}\left(\omega_{2}\right)\right]^{1 / 2} \tag{Eq.92}
\end{equation*}
$$

The Wald $100(1-\alpha / 2) \%$ upper confidence limit on the median of the lower paired effect $\omega_{1}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1+\Delta)$ was derived as

$$
\begin{equation*}
\hat{\mathrm{M}}_{\omega_{1}}+\mathrm{t}_{1-\alpha / 2 ; \mathrm{p}} \operatorname{SE}\left(\hat{\mathrm{M}}_{\omega_{1}}\right) \tag{Eq.93}
\end{equation*}
$$

The Wald $100(1-\alpha / 2) \%$ lower confidence limit on the median of the upper paired effect $\omega_{2}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)$ was derived as

$$
\begin{equation*}
\hat{\mathrm{M}}_{\omega_{2}}-\mathrm{t}_{1-\alpha / 2 ; \mathrm{p}} \mathrm{SE}\left(\hat{\mathrm{M}}_{\omega_{2}}\right) \tag{Eq.94}
\end{equation*}
$$

Equations 93 and 94 are based on the inverse $t$ CDF employed $100(1-\alpha) \%$ confidence with $p=U_{n}-L_{n}-1$ degrees of freedom, where $U_{n}$ and $L_{n}$ were defined for Equation 77 (p. 95). For Equations 91-94, coverage (i.e., observed confidence level) was computed
with 10,000 nonparametric bootstrap simulations for each AE type and $\operatorname{SRS}$ of sizes $\mathrm{n}=$ 30 and 250.

IRT Power Functions. The general form of the IRT power function, based on the non-central F-distribution (Pearson \& Hartley, 1951), was

$$
\begin{equation*}
1-\mathrm{F}_{1-\alpha ; \mathrm{ndf}, \mathrm{ddf}, \mathrm{NCP}}^{-1} \tag{Eq.95}
\end{equation*}
$$

$n d f$ denotes the numerator degrees of freedom (i.e., 1). $d d f$ denotes the denominator degrees of freedom. This equals $n-1$ for paired studies where $n$ is the proposed sample size. $N C P$ is the non-centrality parameter of the F-distribution.

The power function for the superiority hypothesis set was based on the mean or median of the paired effect $\omega=\mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}$. For $n-1$ degrees of freedom, the noncentrality parameter for the mean of the effect was derived as

$$
\begin{equation*}
\mathrm{n}\left[\frac{\overline{\mathrm{X}}_{\omega}^{2}}{\operatorname{MSE}[\omega]}\right] \tag{Eq.96}
\end{equation*}
$$

For $p=U_{n}-L_{n}-1$ degrees of freedom, the non-centrality parameter for the median of the effect was derived as

$$
\begin{equation*}
\mathrm{n}\left[\frac{\mathrm{M}_{\omega}^{2}}{\mathrm{SE}[\hat{\mathrm{M}}]}\right] \tag{Eq.97}
\end{equation*}
$$

The power function for the equivalence hypothesis set was based on the mean or median of the paired effects $\omega_{1}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1+\Delta)$ and $\omega_{2}: \mathrm{A}_{\mathrm{T}}-\mathrm{A}_{\mathrm{R}}(1-\Delta)$. For $n-1$ degrees of freedom, the non-centrality parameter for the mean of the effects was derived as

$$
\begin{equation*}
\mathrm{n}\left[\frac{\bar{X}_{\text {effect }}^{2}}{\operatorname{MSE}[\text { effect }]}\right] \tag{Eq.98}
\end{equation*}
$$

For $p=U_{n}-L_{n}-1$ degrees of freedom, the non-centrality parameter for the median of the effects was derived as

$$
\begin{equation*}
\mathrm{n}\left[\frac{\mathrm{M}_{\omega}^{2}}{\mathrm{SE}[\hat{\mathrm{M}}]}\right] \tag{Eq.99}
\end{equation*}
$$

For estimation based on the mean or median, the sample size for equivalence equaled the maximum sample size $\max \left\{\mathrm{n}_{\omega_{1}}, \mathrm{n}_{\omega_{2}}\right\}$.

## Research Question 3

The last research question is repeated as
How do IRT sample size requirements for superiority and equivalence study objectives compare to the existing classical methods for paired binomial variables?

IRT power functions were derived by the author for both superiority and equivalence study objectives. If the IRT sample sizes were comparable to efficacy requirements, it may be possible to power ICTs for clinically relevant AE hypothses. Sample sizes based on existing methods do not typically result in financially feasible sample size requirements for AE hypotheses. If IRT sample sizes for AE hypotheses are not comparable to efficacy requirements, then IRT may only take a tertiary role as a secondary analytic method.

Existing methods for computing sample size requirements on paired binomial variables for superiority and equivalence study objectives were presented in Chapter II. This study compared sample size requirements from these methods to the IRT models with the best statistical properties investigated under Research Question 1 and 2. This formal comparison was necessary because IRT is not well-known to statisticians in the pharmaceutical industry. In order for these statisticians to accept, promote, adopt, and incorporate the presented IRT methodology into their ICTs, IRT will need to demonstrate a clear scientific and financial advantage over current practice.

## CHAPTER IV

## RESULTS

This chapter presents the results for analyzing rare types of AEs with the proposed IRT methodology. Assessing the distributional forms of the applicable historical data was the first step in this analysis. Next, Monte Carlo (MC) simulations based on the historical incidence of these data were performed and their quality was then assessed. In the next section of this chapter, the assumptions of the three Bayesian IRT models will be investigated with these data. Next, analyses were performed to assess superiority and equivalence study objectives for simulations based on AE historical incidence and four data patterns, which represented various scenarios of AE characteristics. The last section of this chapter will compare sample size requirements for superiority and equivalence study objectives between existing methods and the IRT approaches derived by the author.

Distributional Forms of Historical Data

The relative frequency (RF) statistic was used to determine if the beta, gamma, or normal probability density functions (PDFs) could be used to model the transfusionrelated AEs. Table 7 (p. 106) presents the probabilities for these PDFs, which are visually displayed in Figure 7 (p. 107). The RF equals the expected number of binary events divided by the total number of events reported. Known PDFs that closely follow the empirical distribution of the RFs can be used for modeling the data under investigation.

The RF of the $k=9$ types of transfusion-related AEs investigated in this study was presented in Table 3 (p. 81). The incidence and RF of these events are presented in Figure 6 (p. 107) and Figure 7 (p. 107), respectively. For this analysis, the RF was computed as the expected number of AEs (last column of Table 3, p. 81) divided by the total number of AEs in this column. The probability of being equal to or less than the RFs for each type of AE was then computed with the beta, gamma, and normal PDFs.

As an example, the beta probability for "Delayed Serologic Transfusion Reaction" was computed as

$$
\operatorname{Pr}(\mathrm{X} \leq \mathrm{RF})=\int_{0}^{0.74634} \frac{\Gamma(\mathrm{a}+\mathrm{b})}{\Gamma(\mathrm{a}) \Gamma(\mathrm{b})} \mathrm{x}^{\mathrm{a}-1}(1-\mathrm{x})^{\mathrm{b}-1}=0.83962,
$$

where the shape $(a)$ and location $(b)$ parameters both were set to a value of 2 . For this AE , the gamma probability was computed as

$$
\operatorname{Pr}(\mathrm{X} \leq \mathrm{RF})=\int_{0}^{0.74634} \frac{1}{\Gamma(\mathrm{a}) \mathrm{b}^{\mathrm{a}}} \mathrm{x}^{\mathrm{a}-1} \mathrm{e}^{-\mathrm{x} / \mathrm{b}} \mathrm{dx}=0.52590
$$

where the shape $(a)$ and location $(b)$ parameters both were set to a value of 1 . Last, the normal probability was computed as

$$
\operatorname{Pr}(\mathrm{X} \leq \mathrm{RF})=\int_{-\infty}^{0.74634} \frac{1}{\sigma(2 \pi)^{1 / 2}} \exp \left\{-\frac{(\mathrm{x}-\mu)^{2}}{2 \sigma^{2}}\right\} \mathrm{dx}=0.77227,
$$

for mean 0 and variance 1 .
Table 7 (p. 106) and Figure 7 (p. 107) demonstrate that the beta and gamma PDFs reasonably followed the empirical distribution of the RFs. This finding was not observed for the normal PDF. Irrespective of these findings (Program 1.sas, Appendix M), the
impact of these three PDFs was compared on the discrimination parameter estimates obtained from the 2-PL, 2-PL EX, and 2-PL MEX IRT models for each type of AE.

## Table 7

Relative Frequency of Transfusion-Related Adverse Events

| Type of Adverse Event | Relative <br> Frequency | Probability |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  |  | Gamma | Beta | Normal |
| Delayed Serologic Transfusion Reaction | 0.74634 | 0.52590 | 0.83962 | 0.77227 |
| Allergic Reaction (including anaphylaxis) | 0.09103 | 0.08701 | 0.02335 | 0.53627 |
| Delayed Hemolytic Transfusion Reaction | 0.13570 | 0.12689 | 0.05025 | 0.55397 |
| Infection (Bacterial Contamination) | 0.00259 | 0.00258 | 0.00002 | 0.50103 |
| Febrile Non-hemolytic Transfusion Reaction | 0.01673 | 0.01659 | 0.00083 | 0.50668 |
| Transfusion Associated Circulatory Overload | 0.00082 | 0.00082 | 0.00000 | 0.50033 |
| Transfusion Associated Graft vs Host Disease | 0.00639 | 0.00637 | 0.00012 | 0.50255 |
| Acute Hemolytic Transfusion Reaction | 0.00023 | 0.00023 | 0.00000 | 0.50009 |
| Hypotensive Transfusion Reaction | 0.00017 | 0.00017 | 0.00000 | 0.50007 |

## Quality of Monte Carlo Simulations

The quality of the algorithm for simulating bivariate binomial AE data was evaluated in terms of the deviation between simulated and target AE rates, the correlation between paired treatment groups A and B, and the stability of the AE Predisposition ( $\boldsymbol{\theta})$, discrimination, and difficulty parameters from the 2-PL IRT model. It was assumed that this model represented a worst-case scenario of the three Bayesian IRT models under


Historical Incidence (\%)

Figure 6. Incidence of Transfusion-Related Adverse Events


Historical Incidence (\%)

Figure 7. Relative Frequency of Transfusion-Related Adverse Events
investigation, because of the empirical distribution of the AE historical incidence. Each MC simulation consisted of $N=500,000$ patients for both treatment groups and each type of AE. Smaller sample sizes resulted in algorithms that did not achieve the target AE rates.

Table 8 presents the mean difference (\%) between target and simulated rates and the phi coefficient between the paired treatment groups A and B for each of $k=9$ types of AEs (Program 2.sas). The AE rates of $1,000 \mathrm{MC}$ simulations differed on average at the third decimal point for 4 types of AEs and the second decimal point for the remaining 5 types of AEs. Furthermore, the average phi coefficient between treatments A and B ranged from 0.856 to 0.990 . These findings demonstrated that the MC bivariate binomial simulations resulted in a paired data structure that coincided with the target AE rates.

## Table 8

Metrics of Bivariate Binomial Simulation Algorithm

| Adverse Event | Historical <br> Occurrence | Mean Difference <br> From Target | Mean Phi <br> Coefficient |
| :--- | :---: | :---: | :---: |
| Delayed Serologic Transfusion Reaction | $0.66 \%$ | $-0.00590 \%$ | 0.990 |
| Allergic Reaction (including anaphylaxis) | $0.266 \%$ | $-0.00375 \%$ | 0.984 |
| Delayed Hemolytic Transfusion Reaction | $0.12 \%$ | $-0.00252 \%$ | 0.977 |
| Infection (Bacterial Contamination) | $0.053 \%$ | $-0.00166 \%$ | 0.965 |
| Febrile Non-hemolytic Transfusion Reaction | $0.0489 \%$ | $-0.00160 \%$ | 0.964 |
| Transfusion Associated Circulatory Overload | $0.0168 \%$ | $-0.00097 \%$ | 0.939 |
| Transfusion Associated Graft vs Host Disease | $0.00565 \%$ | $-0.00058 \%$ | 0.892 |
| Acute Hemolytic Transfusion Reaction | $0.004 \%$ | $-0.00047 \%$ | 0.876 |
| Hypotensive Transfusion Reaction | $0.003 \%$ | $-0.00040 \%$ | 0.856 |

The last quality check of the MC simulations pertained to the stability of parameter estimates for each type of AE for Treatment A. MC cumulative assessment plots were constructed for the IRT model parameters AE Predisposition ( $\boldsymbol{\theta})$, discrimination, and difficulty for simple random samples (SRS) of sizes $n=30$ pairs (solid line) and 250 pairs (dashed line). Stability is evidenced with a curve having a zero slope with reasonably small fluctuation or acceptable noise over the $x$-axis. Large fluctuations or spikes over this plane may represent an unstable state.

Figure 8 (p. 110) presents the MC cumulative assessment plots the discrimination parameter for each type of AE (Program 3.sas). The $y$-axis and $x$-axis, respectively, represented the discrimination parameter and number of MCs. This figure demonstrated that the slope of the $x$-plane is zero for each of the $k=9$ types of AEs. For SRS of sizes $n$ $=30$ and 250, the discrimination parameter appeared to stabilize quickly for AE types 2, $3,4,8$, and 9 , and by $1,000 \mathrm{MC}$ simulations for the remaining AE types. Overall, the discrimination parameter varied within a margin of $\pm 0.2$ for all AE types.

Figure 9 (p.110) presents the MC cumulative assessment plots for the difficulty parameter (Program 4.sas). The $y$-axis and $x$-axis, respectively, represented the difficulty parameter and number of MCs. This figure demonstrated that the slope of the $x$-plane is zero for each of the $k=9$ types of AEs. For SRS of sizes $n=30$ and 250, the difficulty parameter stabilized quickly for AE types 8 and 9 , and by $1,000 \mathrm{MC}$ simulations for the remaining AE types. Overall, the difficulty parameter varied within a margin of $\pm 0.5$ for the first type of AE and $\pm 0.3$ for the remaining types of AEs.


Figure 8. 2-PL: Simulation Plots for the Discrimination Parameter


Figure 9. 2-PL: Simulation Plots for the Difficulty Parameter

Figure 10 presents the MC cumulative assessment plot for the AE Predisposition ( $\boldsymbol{\theta})$ parameter across all patients (Program 5.sas). This figure demonstrated that the slope of the $x$-plane is zero. The AE Predisposition ( $\boldsymbol{\theta}$ ) parameter stabilized quickly for the SRS of size $n=250$, and by $1,000 \mathrm{MC}$ simulations for the SRS of size $n=30$. Overall, AE Predisposition $(\boldsymbol{\theta})$ varied within a margin of $\pm 0.05$.


Figure 10. 2-PL: Simulation Plots for the AE Predisposition ( $\boldsymbol{\theta}$ ) Parameter

## Comparison of AUC Approximations

This section compares the linear trapezoid and spline approximations to the exact area under the two Item Response Functions (IRFs) presented in Figure 3 (p. 30). For both of these IRFs, the ability parameter ranged from -2.6 to +2.9 . For IRT Curve 1 or IRF 1, the discrimination $(\boldsymbol{A})$ and difficulty $(\boldsymbol{B})$ parameters, respectively, were arbitrarily
set to -0.5 and -2.0 . For IRT Curve 2 or IRF 2 , the discrimination $(\boldsymbol{A})$ and difficulty $(\boldsymbol{B})$ parameters, respectively, were arbitrarily set to +2.0 and +0.5 .

The exact AUC of IRF 1 was computed as $\int_{-2.6}^{2.9} \frac{\mathrm{e}^{-\mathrm{a}(\theta-\mathrm{b})}}{1+\mathrm{e}^{-\mathrm{a}(\theta-\mathrm{b})}} \mathrm{d} \theta=1.54317$. The linear trapezoid and spline approximations to the exact AUC were respectively computed as 1.54976 and 1.53546 . These approximations respectively deviated by a bias factor of -0.00659 and 0.00771 . When an IRF is relatively linear, the linear trapezoid and spline approximations to the exact area under IRFs are expected to be comparable.

The exact AUC of IRF 2 was computed as 2.40308 . The linear trapezoid and spline approximations to the exact AUC were respectively computed as 2.36794 and 2.41006. These approximations respectively deviated by a bias factor of 0.03515 and -0.00697 . When an IRF is relatively nonlinear (e.g., cubic), the spline approximation is expected to better approximate the exact AUC than the linear trapezoid approximation.

## Research Question 1 Results

The first step in performing IRT analysis consisted of evaluating the Bayesian assumptions for each of the parameter estimates by AE type. These assumptions were investigated for the AE Predisposition ( $\boldsymbol{\theta}$ ), discrimination, and difficulty parameters across 1,000 MCs for SRS of sizes $n=30$ and 250 for only Treatment A. Both treatment groups were simulated to possess the same theoretical data structure.

The first assumption was concerned with whether or not multiple chains of starting values (Table 6, p. 86) converged to the same location for each IRT parameter. After demonstration of this assumption, the stationary states for the mean and standard
deviation (SD)of each parameter were evaluated with trace and cumulative GelmanRubin (G-R) convergence plots. The G-R plots were based on $m=10-1$ cumulative blocks of size 10. After stationary states were identified, lag functions for removing serial autocorrelation of the Gibbs sampler estimates were computed. Last, goodness-of-fit was assessed for each IRT model analyzing data from the identified stationary state after serial autocorrelation removal. These assumptions were described in Chapter II.

## 2-PL IRT Model

Multiple Chains. The chains of starting values used were (1) $\boldsymbol{A}^{(i)}=\mathbf{- 1}$ and $\boldsymbol{\theta}^{(i)}=$ $-\mathbf{1}$, (2) $\boldsymbol{A}^{(i)}=+\mathbf{5}$ and $\boldsymbol{\theta}^{(i)}=\mathbf{0}$, and (3) $\boldsymbol{A}^{(i)}=+\mathbf{4}$, and $\boldsymbol{\theta}^{(i)}=+\mathbf{1}$. The choice of these chains was arbitrary, and these chains were intentionally selected to be far apart to stress the IRT model. Across these chains, the mean of the discrimination parameter ranged from -0.02 to 0.02 for $n=30$ and from -0.01 to 0.05 for $n=250$. The mean of the difficulty parameter ranged from 1.95 to 2.15 for $n=30$ and from 2.5 to 3.0 for $n=250$. The mean of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from -0.002 to -0.001 for $n=30$ and from -0.003 to 0.002 for $n=250$. The $S D$ of the discrimination parameter ranged from 0.01 to 0.02 for $n=30$ and from 0.01 to 0.03 for $n=250$. The $S D$ of the difficulty parameter ranged from 0.01 to 0.28 for $n=30$ and from 0.02 to 0.29 for $n=250$. The $S D$ of the AE Predisposition $(\boldsymbol{\theta})$ parameter was 0.01 for $n=30$ and ranged from 0.002 to 0.003 for $n=$ 250. Because the three starting chains converged to the same location for the three IRT parameters, the three chains of starting values were considered interchangeable. Chain 1 appeared to converge at the slowest rate, and represented the worst-case scenario for this
study. As a result, results from this chain are presented below, and chain 2 and 3 results are presented in Appendix A.

Convergence. Starting values defined for chain 1 were used to construct Figures 11-22 that consisted of trace plots (left side) and G-R plots (right side). Trace plots are iterative, and G-R plots are cumulative. For these plots, convergence can be visualized by the flattening of slopes approaching zero. A limitation of these plots is the occurrence of re-ascending curvature. If the slopes converge to zero and later demonstrate quadrature, the interpretation of convergence may become confounded. This situation, a function of sample size, may occur when the observed PDF deviates from the expected PDF.

Figures 11-16 are based on the mean of the IRT parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination, and difficulty. Figures 17-22 are based on the standard deviation of these parameters (Program 7.sas). Plots based on discrimination (Program 8.sas) and difficulty (Program 9.sas) are presented for each type of AE. Plots for AE Predisposition $(\boldsymbol{\theta})$ are presented across $i$ patients (Program 10.sas). Odd and even figure numbers represent sample sizes of 30 and 250 , respectively.

Mean of Parameter Estimates. Figures 11-12 reveal that the mean of the discrimination parameter was stable for all AE types after approximately 25,000 and 15,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures $13-14$ reveal that the mean of the difficulty parameter was stable for all AE types after approximately 25,000 and 10,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 1516 reveal that the mean of the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter was stable across all patients after approximately 30,000 and 25,000 Gibbs sampler iterations for $n=30$ and 250 , respectively.

Standard Deviation (SD) of Parameter Estimates. Figures 17-18 reveal that the $S D$ of the discrimination parameter was stable for all AE types after approximately 10,000 and 30,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 1920 reveal that the mean of the difficulty parameter was stable for all AE types after approximately 15,000 and 30,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 21-22 reveal that the $S D$ of the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter was stable across all patients after approximately 25,000 and 35,000 Gibbs sampler iterations for $n=30$ and 250 , respectively.


Figure 11. 2-PL: Trace and G-R Plot for Mean Discrimination ( $n=30$ )


Figure 12. 2-PL: Trace and G-R Plot for Mean Discrimination ( $n=250$ )


Figure 13. 2-PL: Trace and G-R Plot for Mean Difficulty ( $n=30$ )


Figure 14. 2-PL: Trace and G-R Plot for Mean Difficulty ( $n=250$ )


Figure 15. 2-PL: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30)$


Figure 16. 2-PL: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=250)$


Figure 17. 2-PL: Trace and G-R Plot for $S D$ Discrimination $(n=30)$


Figure 18. 2-PL: Trace and G-R Plot for $S D$ Discrimination $(n=250)$

| AE1: Blue; AE2: Red; AE3: Green AE4: Cyan; AE5: Orange; AE6: Black AE7: Brown; AE8: Olive; AE9: Gold | AE1: Blue; AE2: Red; AE3: Green AE4: Cyan; AE5: Orange; AE6: Black AE7: Brown; AE8: Olive; AE9: Gold |
| :---: | :---: |
|  |  |

Figure 19. 2-PL: Trace and G-R Plot for $S D$ Difficulty ( $n=30$ )


Figure 20. 2-PL: Trace and G-R Plot for $S D$ Difficulty $(n=250)$


Figure 21. 2-PL: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=30)$


Figure 22. 2-PL: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=250)$

The next step in evaluating the Bayesian IRT models identified lag functions that remove autocorrelation from the stationary states. Lags were investigated for the stationary state that consisted of Gibbs sampler iterations 25,001 to 50,000 for $n=30$ and 250.

Autocorrelation. For the discrimination parameter presented in Table 9 (p. 122), the lags that resulted in nonsignificant autocorrelation for the $k=9$ types of AEs ranged from 115 to 185 for $n=30$ and 230 to 350 for $n=250$ (Program 11.sas). Estimation of this parameter was then based on systematic random samples of lags 185 and 350 for all types of AEs for $n=30$ and 250, respectively. For the difficulty parameter presented in Table 10 (p. 122), the lags that resulted in nonsignificant autocorrelation for the $k=9$ types of AEs ranged from 90 to 145 for $n=30$ and 215 to 350 for $n=250$. Estimation of

## Table 9

2-PL: Autocorrelation of Discrimination by AE Type

| Event | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 1 | -0.0428 | $(-0.0861,0.0005)$ | 120 | -0.0567 | $(-0.1368,0.0234)$ | 230 |
| 2 | -0.0358 | $(-0.0800,0.0084)$ | 135 | -0.0919 | $(-0.1871,0.0034)$ | 255 |
| 3 | -0.0106 | $(-0.0530,0.0318)$ | 155 | -0.0717 | $(-0.1554,0.0119)$ | 260 |
| 4 | -0.0177 | $(-0.0601,0.0247)$ | 115 | -0.0395 | $(-0.1224,0.0434)$ | 305 |
| 5 | -0.0259 | $(-0.0761,0.0242)$ | 185 | -0.0367 | $(-0.1252,0.0517)$ | 250 |
| 6 | -0.0389 | $(-0.0821,0.0044)$ | 120 | -0.0774 | $(-0.1589,0.0041)$ | 315 |
| 7 | -0.0434 | $(-0.0896,0.0029)$ | 135 | -0.0871 | $(-0.1830,0.0087)$ | 350 |
| 8 | -0.0473 | $(-0.0947,0.0001)$ | 120 | -0.0815 | $(-0.1668,0.0038)$ | 230 |
| 9 | -0.0274 | $(-0.0653,0.0106)$ | 125 | -0.0793 | $(-0.1597,0.0012)$ | 280 |

Table 10
2-PL: Autocorrelation of Difficulty by AE Type

| Event | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 1 | -0.0374 | $(-0.0834,0.0087)$ | 100 | -0.0789 | $(-0.1593,0.0015)$ | 215 |
| 2 | -0.0407 | $(-0.0825,0.0011)$ | 90 | -0.0874 | $(-0.1793,0.0045)$ | 350 |
| 3 | -0.0408 | $(-0.0844,0.0029)$ | 145 | -0.0873 | $(-0.1774,0.0028)$ | 345 |
| 4 | -0.0365 | $(-0.0759,0.0029)$ | 105 | -0.0660 | $(-0.1483,0.0163)$ | 265 |
| 5 | -0.0338 | $(-0.0758,0.0081)$ | 90 | -0.1117 | $(-0.2244,0.0010)$ | 305 |
| 6 | -0.0388 | $(-0.0848,0.0072)$ | 135 | -0.0838 | $(-0.1695,0.0020)$ | 290 |
| 7 | -0.0372 | $(-0.0805,0.0060)$ | 110 | -0.0647 | $(-0.1673,0.0378)$ | 275 |
| 8 | -0.0224 | $(-0.0639,0.0191)$ | 115 | -0.0851 | $(-0.1767,0.0064)$ | 250 |
| 9 | -0.0453 | $(-0.0908,0.0002)$ | 100 | -0.0840 | $(-0.1731,0.0051)$ | 315 |

this parameter was then based on systematic random samples of lags 145 and 350 for all types of AEs for $n=30$ and 250, respectively. Last, systematic random samples of lags 330 for $n=30$ and 565 for $n=250$ were used to remove the serial autocorrelation from the AE Predisposition $(\boldsymbol{\theta})$ parameter across all patients (Table 11).

Table 11
2-PL: Autocorrelation of AE Predisposition ( $\boldsymbol{\theta}$ )

| SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 0.0007 | $(-0.0040,0.0055)$ | 330 | -0.0011 | $(-0.0024,0.0002)$ | 565 |

Goodness-of-Fit. Bock's Index (BI), posterior probability interval (PPI) plots, and posterior residual plots were used to evaluate the goodness-of-fit (GoF) of each type of AE. The mean of the BI statistic ranged from 22.1 to 27.5 for $n=30$ and 162.6 to 249.5 for $n=250$. These results exceeded the critical value $\chi_{\mathrm{S}-\mathrm{m}, 0.1}^{2}=\chi_{6,0.1}^{2}=2.204$ for block size $S=10$ and $m=4$ IRT model parameters. These parameters were the latent trait $\boldsymbol{Z}$, AE Predisposition $(\boldsymbol{\theta})$, discrimination $\boldsymbol{A}$, and difficulty $\boldsymbol{B}$. This result implied that the 2-PL IRT model poorly fit the data for all AE types. In addition to the concerns of Toribio (2006) for using the BI statistic for gauging the GoF of Bayesian IRT models, using this statistic was further exacerbated because the expected values of the off-diagonals or discordant pairs were simulated to be less than 5 . As a result, IRT model fit should not be decided solely on the BI statistic, and assisted with visual displays.

Evidence of acceptable fit was visualized by plots that exhibited patient discriminatory power, monotonicity, the observed data points were contained within the Posterior Probability Interval (PPI), and the residuals were small. Figures 23-24 (p. 125) comprise the PPI plots (left side) and posterior residual plots (right side) for AE type 1 for $n=30$ (Program 12.sas) and $n=250$ (Program 13.sas), respectively. The plots on the left were used to visualize patient discriminatory power. Given a monotinic IRT, the greater the curvature (i.e., large slopes), the greater the IRF was able to discriminate patients on AE Predisposition ( $\boldsymbol{\theta})$. Plots on the right were used to determine how observed values compared to predicted values for each patient. As the residual decreased in size, the reliability of the IRT model increased.

Patient discriminatory power was not exhibited in these plots. Second, the observed data points were contained within the PPIs. Next, the posterior residual plots revealed that the residuals were small. The same conclusions were reached for the remaining types of AEs for both $n=30$ and 250. Given that patient discriminatory power was not present in the GoF plots in addition to the BI statistic findings, the fit of the 2-PL IRT model was considered insufficient. Poor fit was evidenced by the observation that the empirical distribution of the IRFs did not follow a logistic CDF, and this CDF was not always monotonic. Good fit would follow this density with greater vertical spread of the probability of AE Predisposition $(\boldsymbol{\theta})$. This conclusion was consistent for the remaining types of AEs for both $n=30$ and 250 (Appendix D).


Figure 23. 2-PL: Bayesian PPI and Residual Plot for AE $1(n=30)$


Figure 24. 2-PL: Bayesian PPI and Residual Plot for AE $1(n=250)$

Parameter Estimates. After all Bayesian IRT model assumptions were assessed, estimators of the IRT parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination $(\boldsymbol{A})$, and difficulty $(\boldsymbol{B})$ were computed. Table 12 demonstrates that the mean $(S D)$ of the discrimination parameter ranged from $-0.07(0.556)$ to $0.11(0.547)$ for $n=30$ and $-0.06(0.359)$ to 0.08 (0.341) for $n=250$. Table 13 (p. 127) demonstrates that the mean $(S D)$ of the difficulty parameter ranged from $1.97(0.467)$ to $2.19(0.551)$ for $n=30$ and $2.46(0.311)$ to 3.06 (0.496) for $n=250$. Due to space requirements, summary statistics for the AE Predisposition ( $\boldsymbol{\theta})$ parameter are presented in Appendix E. Tables E-1 and E-2 demonstrate that the mean $(S D)$ of this parameter ranged from -0.02 (0.019) to 0.29 (0.016) for $n=30$ and $-0.38(0.007)$ to $0.41(0.026)$ for $n=$ size 250 .

Table 12
2-PL: Discrimination by AE Type

| AE | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean |
| 1 | $-0.07(0.556)$ | -0.10 | $(-0.16,0.03)$ | $0.07(0.458)$ | 0.07 | $(-0.03,0.18)$ |
| 2 | $0.06(0.490)$ | 0.05 | $(-0.02,0.14)$ | $0.08(0.341)$ | 0.10 | $(-0.00,0.16)$ |
| 3 | $0.02(0.500)$ | 0.01 | $(-0.06,0.11)$ | $-0.06(0.359)$ | -0.06 | $(-0.14,0.03)$ |
| 4 | $-0.03(0.678)$ | 0.03 | $(-0.14,0.09)$ | $-0.01(0.370)$ | -0.02 | $(-0.10,0.08)$ |
| 5 | $-0.01(0.536)$ | -0.01 | $(-0.11,0.08)$ | $-0.03(0.380)$ | -0.06 | $(-0.12,0.06)$ |
| 6 | $-0.06(0.543)$ | -0.05 | $(-0.15,0.04)$ | $0.04(0.368)$ | 0.07 | $(-0.04,0.13)$ |
| 7 | $0.11(0.547)$ | 0.11 | $(0.02,0.20)$ | $0.07(0.329)$ | 0.09 | $(-0.01,0.14)$ |
| 8 | $-0.01(0.513)$ | 0.04 | $(-0.09,0.08)$ | $0.07(0.422)$ | -0.01 | $(-0.03,0.17)$ |
| 9 | $-0.04(0.553)$ | -0.01 | $(-0.13,0.06)$ | $-0.04(0.398)$ | -0.02 | $(-0.13,0.05)$ |

Table 13
2-PL: Difficulty by AE Type

| AE | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean |
| 1 | $2.03(0.525)$ | 1.95 | $(1.95,2.11)$ | $2.46(0.311)$ | 2.43 | $(2.38,2.53)$ |
| 2 | $2.18(0.512)$ | 2.17 | $(2.10,2.26)$ | $2.85(0.426)$ | 2.80 | $(2.75,2.96)$ |
| 3 | $2.19(0.551)$ | 2.17 | $(2.11,2.27)$ | $2.76(0.460)$ | 2.71 | $(2.66,2.87)$ |
| 4 | $1.97(0.467)$ | 1.93 | $(1.90,2.04)$ | $2.98(0.496)$ | 2.91 | $(2.86,3.09)$ |
| 5 | $2.11(0.560)$ | 2.13 | $(2.02,2.19)$ | $2.90(0.402)$ | 2.84 | $(2.81,3.00)$ |
| 6 | $2.16(0.549)$ | 2.07 | $(2.08,2.24)$ | $2.95(0.422)$ | 2.87 | $(2.85,3.05)$ |
| 7 | $2.14(0.499)$ | 2.09 | $(2.06,2.21)$ | $2.92(0.406)$ | 2.88 | $(2.82,3.01)$ |
| 8 | $2.13(0.508)$ | 2.10 | $(2.05,2.21)$ | $3.06(0.496)$ | 3.03 | $(2.94,3.18)$ |
| 9 | $2.11(0.532)$ | 2.04 | $(2.03,2.19)$ | $3.00(0.398)$ | 3.02 | $(2.90,3.09)$ |

## 2-PL EX IRT Model

Multiple Chains. Across the three chains of starting values, the mean of the discrimination parameter ranged from 0.98 to 0.99 for $n=30$ and from 0.77 to 0.83 for $n=250$. The mean of the difficulty parameter ranged from 1.73 to 1.93 for $n=30$ and from 2.77 to 3.21 for $n=250$. The mean of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from -0.56 to -0.55 for $n=30$ and -0.09 to 0.00 for $n=250$. The $S D$ of the discrimination parameter ranged from 0.002 to 0.013 for $n=30$ and from 0.01 to 0.03 for $n=250$. The $S D$ of the difficulty parameter ranged from 0.01 to 0.32 for $n=30$ and from 0.02 to 0.25 for $n=250$. The $S D$ of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from 0.003 to 0.009 for $n=30$ and from 0.001 to 0.002 for $n=250$. Because the three starting chains converged to the same location for the three IRT parameters, the chains of starting
values were considered interchangeable. As found with the 2-PL IRT model, the first chain appeared to have the slowest rate of convergence. As a result, chain 1 results are presented below, and chain 2 and 3 results are presented in Appendix B.

Convergence. Starting values defined for chain 1 were used to construct Figures $25-36$ that consist of trace plots (left side) and G-R plots (right side). Figures $25-30$ are based on the mean of the IRT parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination, and difficulty. Figures 31-36 are based on the standard deviation of these parameters (Program 14.sas). Plots based on discrimination (Program 15.sas) and difficulty (Program 16.sas) are presented for each type of AE. Plots for AE Predisposition $(\boldsymbol{\theta})$ are presented across $n$ patients (Program 17.sas). Odd and even figure numbers represent sample sizes of 30 and 250 , respectively.

Mean of Parameter Estimates. Figures 25-26 reveal that the mean of the discrimination parameter was stable for all AE types after approximately 15,000 and 25,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures $27-28$ reveal that the mean of the difficulty parameter was stable for all AE types after approximately 15,000 and 10,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 2930 reveal that the mean of the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter was stable across all patients after approximately 30,000 and 15,000 Gibbs sampler iterations for $n=30$ and 250 , respectively.


Figure 25. 2-PL EX: Trace and G-R Plot for Mean Discrimination ( $n=30$ )


Figure 26. 2-PL EX: Trace and G-R Plot for Mean Discrimination ( $n=250$ )


Figure 27. 2-PL EX: Trace and G-R Plot for Mean Difficulty ( $n=30$ )


Figure 28. 2-PL EX: Trace and G-R Plot for Mean Difficulty ( $n=250$ )


Figure 29. 2-PL EX: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30)$


Figure 30. 2-PL EX: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=250)$

Standard Deviation (SD) of Parameter Estimates. Figures 31-32 reveal that the $S D$ of the discrimination parameter was stable for all AE types after approximately 20,000 Gibbs sampler iterations for $n=30$ and 250. Figures $33-34$ reveal that the $S D$ of the difficulty parameter was stable for all AE types after approximately 10,000 and 15,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures $35-36$ reveal that the $S D$ of the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter was stable across all patients after approximately 20,000 Gibbs sampler iterations for $n=30$ and 250 .

Autocorrelation. Lags were investigated for the stationary state that consisted of Gibbs sampler iterations 15,001 to 50,000 for $n=30$ and 250 . For the discrimination parameter presented in Table 14 (p. 136), the lags that resulted in nonsignificant autocorrelation for the $k=9$ types of AEs ranged from 80 to 95 for $n=30$ and 270 to 365 for $n=250$ (Program 18.sas). Estimation of this parameter was then based on systematic random samples of lags 95 and 365 for all types of AEs for $n=30$ and 250, respectively. For the difficulty parameter presented in Table 15 (p. 136), the lags that resulted in nonsignificant autocorrelation for the $k=9$ types of AEs ranged from 155 to 210 for $n=30$ and 370 to 430 for $n=250$. Estimation of this parameter was then based on systematic random samples of lags 210 and 430 for all types of AEs for $n=30$ and 250, respectively. Last, systematic random samples of lags 180 for $n=30$ and 695 for $n=250$ were used to remove the serial autocorrelation from the AE Predisposition $(\boldsymbol{\theta})$ parameter across all patients (Table 16, p. 137).


Figure 31. 2-PL EX: Trace and G-R Plot for $S D$ Discrimination $(n=30)$


Figure 32. 2-PL EX: Trace and G-R Plot for $S D$ Discrimination ( $n=250$ )


Figure 33. 2-PL EX: Trace and G-R Plot for $S D$ Difficulty $(n=30)$

| AE1: Blue; AE2: Red; AE3: Green <br> AE4: Cyan; AE5: Orange; AE6: Black <br> AE7: Brown; AE8: Olive; AE9: Gold | AE1: Blue; AE2: Red; AE3: Green <br> AE4: Cyan; AE5: Orange; AE6: Black <br> AE7: Brown; AE8: Olive; AE9: Gold |
| :---: | :---: |
|  |  |

Figure 34. 2-PL EX: Trace and G-R Plot for $S D$ Difficulty ( $n=250$ )


Figure 35. 2-PL EX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=30)$


Figure 36. 2-PL EX: Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(n=250)$

## Table 14

2-PL EX: Autocorrelation of Discrimination by AE Type

|  | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Event | Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 1 | -0.0159 | $(-0.0346,0.0028)$ | 95 | -0.0538 | $(-0.1140,0.0064)$ | 270 |
| 2 | -0.0176 | $(-0.0370,0.0018)$ | 85 | -0.0526 | $(-0.1063,0.0012)$ | 315 |
| 3 | -0.0155 | $(-0.0342,0.0032)$ | 85 | -0.0438 | $(-0.0994,0.0118)$ | 305 |
| 4 | -0.0175 | $(-0.0362,0.0013)$ | 80 | -0.0368 | $(-0.0881,0.0146)$ | 295 |
| 5 | -0.0184 | $(-0.0371,0.0004)$ | 85 | -0.0433 | $(-0.0945,0.0080)$ | 330 |
| 6 | -0.0179 | $(-0.0363,0.0004)$ | 90 | -0.0500 | $(-0.1043,0.0043)$ | 300 |
| 7 | -0.0186 | $(-0.0375,0.0003)$ | 85 | -0.0460 | $(-0.0956,0.0037)$ | 365 |
| 8 | -0.0195 | $(-0.0395,0.0005)$ | 85 | -0.0498 | $(-0.1027,0.0032)$ | 330 |
| 9 | -0.0173 | $(-0.0366,0.0020)$ | 90 | -0.0448 | $(-0.0982,0.0085)$ | 340 |

## Table 15

2-PL EX: Autocorrelation of Difficulty by AE Type

| Event | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 1 | -0.0368 | $(-0.0757,0.0021)$ | 155 | -0.0768 | $(-0.1537,0.0001)$ | 370 |
| 2 | -0.0310 | $(-0.0707,0.0086)$ | 170 | -0.0549 | $(-0.1389,0.0292)$ | 410 |
| 3 | -0.0381 | $(-0.0777,0.0015)$ | 185 | -0.0648 | $(-0.1426,0.0130)$ | 405 |
| 4 | -0.0379 | $(-0.0779,0.0021)$ | 210 | -0.0838 | $(-0.1725,0.0049)$ | 385 |
| 5 | -0.0302 | $(-0.0681,0.0077)$ | 185 | -0.0773 | $(-0.1620,0.0075)$ | 400 |
| 6 | -0.0387 | $(-0.0804,0.0031)$ | 195 | -0.0851 | $(-0.1750,0.0048)$ | 430 |
| 7 | -0.0313 | $(-0.0692,0.0066)$ | 195 | -0.0855 | $(-0.1715,0.0004)$ | 400 |
| 8 | -0.0375 | $(-0.0774,0.0023)$ | 180 | -0.0723 | $(-0.1535,0.0089)$ | 420 |
| 9 | -0.0364 | $(-0.0789,0.0062)$ | 205 | -0.0741 | $(-0.1635,0.0152)$ | 380 |

Table 16
2-PL EX: Autocorrelation of AE Predisposition ( $\boldsymbol{\theta}$ )

| SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| -0.0022 | $(-0.0126,0.0083)$ | 180 | -0.0116 | $(-0.0234,0.0002)$ | 695 |

Goodness-of-Fit. Bock's Index (BI), posterior probability interval (PPI) plots, and posterior residual plots were used to evaluate the goodness-of-fit (GoF) of each type of AE. The mean of the BI statistic ranged from 297.4 to 323.0 for $n=30$ and 1832.7 to 1986.1 for $n=250$. These results exceeded the critical value $\chi_{\mathrm{S}-\mathrm{m}, 0.1}^{2}=\chi_{6,0.1}^{2}=2.204$ for block size $S=10$ and $m=4$ IRT model parameters. These parameters were the latent trait $\boldsymbol{Z}$, AE Predisposition $(\boldsymbol{\theta})$, discrimination ( $\boldsymbol{A}$ ), and difficulty $(\boldsymbol{B})$. This result implied that the 2-PL EX IRT model poorly fit the data for all AE types. As previously stated, IRT model fit should not be decided solely on the BI statistic, and assisted with visual displays.

Figure 37 (p. 138) and Figure 38 (p. 139) comprise the PPI plots (left side) and posterior residual plots (right side) for AE type 1 for $n=30$ (Program 19.sas) and $n=250$ (Program 20.sas), respectively. Patient discriminatory power was exhibited in these monotonic plots. Second, the observed data points were contained within the Posterior Probability Intervals (PPIs). Next, the posterior residual plots revealed that the residuals were small. The same conclusions were reached for the remaining types of AEs for both $n=30$ and 250. Although the BI statistic findings imply poor fit, the remaining GoF assumptions were reasonably achieved. As a result, the fit of the 2-PL EX IRT model was
considered acceptable for modeling rare binary event data. This conclusion was consistent for the remaining types of AEs for both $n=30$ and 250 (Appendix F).


Figure 37. 2-PL EX: Bayesian PPI and Residual Plot for AE $1(n=30)$

Parameter Estimates. After all Bayesian IRT model assumptions were assessed, estimators of the IRT parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination $(\boldsymbol{A})$, and difficulty (B) were computed. Table 17 (p. 140) demonstrates that the mean $(S D)$ of the discrimination parameter ranged from $0.97(0.238)$ to $0.99(0.250)$ for $n=30$ and 0.77 ( 0.220 ) to $0.82(0.226)$ for $n=250$. Table 18 (p. 140) demonstrates that the mean (SD) of the difficulty parameter ranged from $1.82(0.495)$ to $1.94(0.554)$ for $n=30$ and 2.82 ( 0.356 ) to 3.21 ( 0.433 ) for $n=250$. Due to space requirements, summary statistics for the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter are presented in Appendix G. Tables G-1 and G-2
demonstrate that the mean $(S D)$ of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from $-0.63(0.044)$ to $-0.38(0.358)$ for $n=30$ and $-0.16(0.133)$ to $\approx 0.00(0.320)$ for $n=250$.


Figure 38. 2-PL EX: Bayesian PPI and Residual Plot for AE $1(n=250)$

## 2-PL MEX IRT Model

Multiple Chains. Across the three chains of starting values, the mean of the discrimination parameter ranged from 2.1 to 2.4 for $n=30$ and from -0.01 to 0.10 for $n=250$. The mean of the difficulty parameter ranged from 1.30 to 1.45 for $n=30$ and from 2.45 to 3.00 for $n=250$. The mean of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from -0.87 to -0.65 for $n=30$ and from -0.005 to -0.003 for $n=250$. The $S D$ of the discrimination parameter ranged from 0.1 to 0.3 for $n=30$ and from 0.02 to 0.10 for

Table 17
2-PL EX: Discrimination by AE Type

| AE | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean |
| 1 | $0.99(0.245)$ | 0.99 | $(0.97,1.02)$ | $0.82(0.226)$ | 0.82 | $(0.77,0.86)$ |
| 2 | $0.98(0.239)$ | 0.97 | $(0.95,1.00)$ | $0.79(0.222)$ | 0.78 | $(0.74,0.83)$ |
| 3 | $0.98(0.243)$ | 0.98 | $(0.96,1.01)$ | $0.77(0.225)$ | 0.78 | $(0.73,0.82)$ |
| 4 | $0.97(0.238)$ | 0.97 | $(0.95,0.99)$ | $0.77(0.220)$ | 0.78 | $(0.73,0.82)$ |
| 5 | $0.98(0.239)$ | 0.98 | $(0.96,1.01)$ | $0.77(0.227)$ | 0.77 | $(0.73,0.82)$ |
| 6 | $0.98(0.247)$ | 0.98 | $(0.96,1.01)$ | $0.77(0.221)$ | 0.76 | $(0.72,0.81)$ |
| 7 | $0.99(0.245)$ | 0.98 | $(0.96,1.01)$ | $0.77(0.222)$ | 0.77 | $(0.72,0.81)$ |
| 8 | $0.98(0.252)$ | 0.98 | $(0.95,1.00)$ | $0.77(0.223)$ | 0.77 | $(0.73,0.82)$ |
| 9 | $0.99(0.250)$ | 0.99 | $(0.96,1.02)$ | $0.77(0.222)$ | 0.77 | $(0.72,0.81)$ |

Table 18
2-PL EX: Difficulty by AE Type

| AE | SRS of Size 30 |  | SRS of Size 250 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean |
| 1 | $1.82(0.495)$ | 1.78 | $(1.75,1.90)$ | $2.82(0.356)$ | 2.80 | $(2.74,2.90)$ |
| 2 | $1.90(0.532)$ | 1.88 | $(1.82,1.98)$ | $3.05(0.403)$ | 3.02 | $(2.96,3.14)$ |
| 3 | $1.92(0.536)$ | 1.89 | $(1.83,2.00)$ | $3.11(0.415)$ | 3.09 | $(3.02,3.20)$ |
| 4 | $1.89(0.548)$ | 1.85 | $(1.80,1.97)$ | $3.18(0.435)$ | 3.15 | $(3.09,3.28)$ |
| 5 | $1.94(0.513)$ | 1.90 | $(1.86,2.01)$ | $3.15(0.427)$ | 3.13 | $(3.06,3.25)$ |
| 6 | $1.89(0.538)$ | 1.86 | $(1.81,1.98)$ | $3.18(0.432)$ | 3.14 | $(3.08,3.27)$ |
| 7 | $1.91(0.538)$ | 1.88 | $(1.83,1.99)$ | $3.20(0.434)$ | 3.16 | $(3.10,3.29)$ |
| 8 | $1.94(0.554)$ | 1.86 | $(1.85,2.02)$ | $3.20(0.441)$ | 3.16 | $(3.11,3.30)$ |
| 9 | $1.92(0.533)$ | 1.90 | $(1.84,2.00)$ | $3.21(0.433)$ | 3.17 | $(3.11,3.30)$ |

$n=250$. The $S D$ of the difficulty parameter ranged from 0.03 to 0.23 for $n=30$ and from 0.025 to 0.24 for $n=250$. The $S D$ of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from 0.025 to 0.35 for $n=30$ and from 0.0015 to 0.0025 for $n=250$. Because the three starting chains converged to the same location for the three IRT parameters, the chains of starting values were considered interchangeable. Following the conclusions of the 2-PL and 2-PL EX IRT models, chain 1 presented the slowest convergence rate. As a result, results for this chain were presented below, and chain 2 and 3 results are presented in Appendix C.

Convergence. Starting values defined for chain 1 were used to construct Figures 39-50 that consist of trace plots (left side) and G-R plots (right side). Figures 39-44 are based on the mean of the IRT parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination, and difficulty. Figures 45-50 are based on the standard deviation of these parameters (Program 21.sas). Plots based on discrimination (Program 22.sas) and difficulty (Program 23.sas) are presented for each type of AE. Plots for AE Predisposition $(\boldsymbol{\theta})$ are presented across $n$ patients (Program 24.sas). Odd and even figure numbers represent sample sizes 30 and 250 , respectively.

Mean of Parameter Estimates. Figures 39-40 reveal that the mean of the discrimination parameter was stable for all AE types after approximately 7,000 and 6,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 41-42 reveal that the mean of the difficulty parameter was stable for all AE types after approximately 7,000 and 6,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 43 and 44 (p. 144) reveal that the mean of the AE Predisposition $(\boldsymbol{\theta})$ across all patients was not stable for both $n=30$ and 250 .


Figure 39. 2-PL MEX: Trace and G-R Plot for Mean Discrimination ( $n=30$ )


Figure 40. 2-PL MEX: Trace and G-R Plot for Mean Discrimination ( $n=250$ )


Figure 41. 2-PL MEX: Trace and G-R Plot for Mean Difficulty $(n=30)$


Figure 42. 2-PL MEX: Trace and G-R Plot for Mean Difficulty ( $n=250$ )


Figure 43. 2-PL MEX: Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30)$


Figure 44. 2-PL MEX: Trace and G-R Plot for Mean AE Predisposition $\boldsymbol{\theta} \boldsymbol{\theta})(n=250)$

Standard Deviation (SD) of Parameter Estimates. Figures 45-46 reveal that the $S D$ of the discrimination parameter was stable for all AE types except AE type 1 after approximately 6,000 and 7,000 Gibbs sampler iterations for $n=30$ and 250, respectively. Figures 47-48 reveal that the $S D$ of the difficulty parameter was stable for all AE types after approximately 7,000 Gibbs sampler iterations for $n=30$. For $n=250$, this stability was not observed for the majority of AE types. Figures 49-50 reveal that the $S D$ of the AE Predisposition ( $\boldsymbol{\theta})$ parameter across all patients was not stable for both $n=30$ and 250.

Autocorrelation. Lags were investigated for the stationary state that consisted of Gibbs sampler iterations 6,001 to 12,000 for $n=30$ and 250 . For the discrimination parameter presented in Table 19 (p. 149), the lags that resulted in nonsignificant autocorrelation for the $k=9$ types of AEs ranged from 155 to 215 for $n=30$ and 140 to 235 for $n=250$ (Program 25.sas). Estimation of this parameter was based on systematic random samples of lags 215 and 235 for all types of AEs for $n=30$ and 250, respectively. For the difficulty parameter presented in Table 20 (p. 149), the lags that resulted in nonsignificant autocorrelation for the $k=9$ types of AEs ranged from 50 to 100 for $n=$ 30 and 130 to 205 for $n=250$. Estimation of this parameter was based on systematic random samples of lags 100 and 205 for $n=30$ and 250 , respectively. Last, systematic random sample of lags 310 for $n=30$ and 90 for $n=250$ were used to remove the serial autocorrelation from the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter across all patients (Table 21, p. 150).


Figure 45. 2-PL MEX: Trace and G-R Plot for $S D$ Discrimination ( $n=30$ )


Figure 46. 2-PL MEX: Trace and G-R Plot for $S D$ Discrimination ( $n=250$ )


Figure 47. 2-PL MEX: Trace and G-R Plot for $S D$ Difficulty ( $n=30$ )


Figure 48. 2-PL MEX: Trace and G-R Plot for SD Difficulty ( $n=250$ )


Figure 49. 2-PL MEX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=30)$


Figure 50. 2-PL MEX: Trace and G-R Plot for $S D$ AE Predisposition $(\boldsymbol{\theta})(n=250)$

## Table 19

2-PL MEX: Autocorrelation of Discrimination by AE Type

| Event | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 1 | -0.1796 | $(-0.3653,0.0060)$ | 180 | -0.1853 | $(-0.3896,0.0189)$ | 190 |
| 2 | -0.1603 | $(-0.3435,0.0230)$ | 190 | -0.1450 | $(-0.3014,0.0114)$ | 175 |
| 3 | -0.1667 | $(-0.3576,0.0243)$ | 215 | -0.1781 | $(-0.3614,0.0053)$ | 235 |
| 4 | -0.1478 | $(-0.3412,0.0455)$ | 185 | -0.1402 | $(-0.3105,0.0301)$ | 145 |
| 5 | -0.1680 | $(-0.3615,0.0255)$ | 215 | -0.1020 | $(-0.3469,0.1429)$ | 190 |
| 6 | -0.1733 | $(-0.3873,0.0406)$ | 205 | -0.1293 | $(-0.3092,0.0505)$ | 150 |
| 7 | -0.1353 | $(-0.3008,0.0302)$ | 155 | -0.1183 | $(-0.2942,0.0576)$ | 165 |
| 8 | -0.1762 | $(-0.3685,0.0161)$ | 205 | -0.1509 | $(-0.3209,0.0191)$ | 140 |
| 9 | -0.1846 | $(-0.3713,0.0021)$ | 210 | -0.1598 | $(-0.3653,0.0457)$ | 175 |

Table 20
2-PL MEX: Autocorrelation of Difficulty by AE Type

| Event | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| 1 | -0.0809 | $(-0.1621,0.0003)$ | 80 | -0.1565 | $(-0.3218,0.0089)$ | 155 |
| 2 | -0.0577 | $(-0.1383,0.0230)$ | 80 | -0.1719 | $(-0.3447,0.0009)$ | 145 |
| 3 | -0.0772 | $(-0.1577,0.0033)$ | 60 | -0.1557 | $(-0.3314,0.0201)$ | 130 |
| 4 | -0.0562 | $(-0.1289,0.0164)$ | 60 | -0.1196 | $(-0.3042,0.0650)$ | 205 |
| 5 | -0.0625 | $(-0.1443,0.0193)$ | 80 | -0.1915 | $(-0.3841,0.0012)$ | 160 |
| 6 | -0.0562 | $(-0.1347,0.0223)$ | 100 | -0.1494 | $(-0.3359,0.0371)$ | 180 |
| 7 | -0.0713 | $(-0.1441,0.0015)$ | 50 | -0.1571 | $(-0.3666,0.0523)$ | 160 |
| 8 | -0.0804 | $(-0.1638,0.0030)$ | 70 | -0.1386 | $(-0.3168,0.0396)$ | 160 |
| 9 | -0.0759 | $(-0.1577,0.0058)$ | 95 | -0.1730 | $(-0.3714,0.0254)$ | 160 |

Table 21
2-PL MEX: Autocorrelation of AE Predisposition ( $\boldsymbol{\theta}$ )

| SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Mean ACF | 95\% CI (2-sided) | Lag | Mean ACF | 95\% CI (2-sided) | Lag |
| -0.0436 | $(-0.0910,0.0038)$ | 310 | 0.0111 | $(-0.0059,0.0282)$ | 90 |

Goodness-of-Fit. Bock's Index (BI), posterior probability interval (PPI) plots, and posterior residual plots were used to evaluate the goodness-of-fit (GoF) of each type of AE. The mean of the BI statistic ranged from 647.3 to 1095.0 for $n=30$ and 190.7 to 321.9 for $n=250$. These results exceeded the critical value $\chi_{\mathrm{S}-\mathrm{m}, 0.1}^{2}=\chi_{6,0.1}^{2}=2.204$ for block size $S=10$ and $m=4$ IRT model parameters. These parameters were the latent trait $\boldsymbol{Z}$, AE Predisposition $(\boldsymbol{\theta})$, discrimination $(\boldsymbol{A})$, and difficulty $(\boldsymbol{B})$. This result implied that the 2-PL MEX IRT model poorly fit the data for all AE types. As previously stated, IRT model fit should not be decided solely on the BI statistic, and assisted with visual displays.

Figures 51-52 comprise the PPI plots (left side) and posterior residual plots (right side) for AE type 1 for $n=30$ (Program 26.sas) and $n=250$ (Program 27.sas), respectively. Patient discriminatory power was and was not exhibited for $n=30$ and 250, respectively. Second, the observed data points were contained within the Posterior Probability Intervals (PPIs). Last, the posterior residual plots revealed that the residuals were small. The same conclusions were reached for the remaining types of AEs for both $n$ $=30$ and 250. Given that patient discriminatory power and monotonicity were not consistently exhibited in the GoF plots in addition to the BI statistic findings, the fit of the

2-PL MEX IRT model was considered unacceptable. This conclusion was consistent for the remaining types of AEs for both $n=30$ and 250 (Appendix H).

Parameter Estimates. After all Bayesian IRT model assumptions were assessed, estimators of the IRT parameters AE Predisposition $(\boldsymbol{\theta})$, discrimination $(\boldsymbol{A})$, and difficulty (B) were computed. Table 22 (p. 153) demonstrates that the mean $(S D)$ of the discrimination parameter ranged from 1.73 (1.150) to 2.07 (1.288) for $n=30$ and -0.04 ( 0.373 ) to $0.12(0.368)$ for $n=250$. Table 23 (p.153) demonstrates that the mean $(S D)$ of the difficulty parameter ranged from $1.02(0.574)$ to $1.20(0.559)$ for $n=30$ and 2.56 ( 0.377 ) to 3.02 ( 0.441 ) for $n=250$. Due to space requirements, summary statistics for the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter are presented in Appendix I. Tables I-1 and I-2 demonstrate that the mean $(S D)$ of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from $-0.79(0.446)$ to $-0.50(0.487)$ for $n=30$ and $-0.45(0.291)$ to $0.40(0.395)$ for $n=250$.

Summary of IRT Model Assumptions. This section evaluates the assumptions of the three Bayesian IRT models for rare binomial events. For the 2-PL IRT model, the mean of the discrimination parameter ranged from -0.07 to 0.11 and -0.06 to 0.08 for $n=30$ and 250 , respectively. The mean of the difficulty parameter ranged from 1.97 to 2.19 and 2.46 to 3.06 for $n=30$ and 250 , respectively. The mean of the AE Predisposition ( $\boldsymbol{\theta})$ parameter ranged from -0.02 to 0.29 and -0.38 to 0.41 for $n=30$ and 250 , respectively. Trace and G-R plots for the three chains of starting values demonstrated consistent convergence of the mean and standard deviation of the discrimination, difficulty, and AE Predisposition $(\boldsymbol{\theta})$ IRT parameters. Last, irrespective of the posterior probability interval and residual plots, the 2-PL IRT model did not exhibit


Figure 51. 2-PL MEX: Bayesian PPI and Residual Plot for AE 1 ( $n=30$ )


Figure 52. 2-PL MEX: Bayesian PPI and Residual Plot for AE 1 ( $n=250$ )

Table 22
2-PL MEX: Discrimination by AE Type

| $\mathbf{A E}$ | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean |
| 1 | $1.87(1.218)$ | 1.66 | $(1.71,2.97)$ | $0.03(0.412)$ | -0.01 | $(-0.14,0.21)$ |
| 2 | $1.79(1.209)$ | 1.61 | $(1.62,2.87)$ | $0.04(0.384)$ | 0.05 | $(-0.12,0.21)$ |
| 3 | $1.73(1.150)$ | 1.59 | $(1.57,2.75)$ | $0.08(0.393)$ | 0.11 | $(-0.09,0.25)$ |
| 4 | $1.95(1.196)$ | 1.80 | $(1.82,3.06)$ | $-0.04(0.373)$ | -0.05 | $(-0.20,0.12)$ |
| 5 | $1.86(1.112)$ | 1.82 | $(1.75,2.90)$ | $0.12(0.368)$ | 0.15 | $(-0.04,0.28)$ |
| 6 | $1.88(1.126)$ | 1.80 | $(1.76,2.92)$ | $0.04(0.385)$ | 0.07 | $(-0.13,0.21)$ |
| 7 | $2.07(1.288)$ | 1.97 | $(1.92,3.25)$ | $0.08(0.378)$ | 0.05 | $(-0.08,0.24)$ |
| 8 | $1.90(1.062)$ | 1.76 | $(1.83,2.92)$ | $0.02(0.411)$ | -0.03 | $(-0.15,0.20)$ |
| 9 | $1.94(1.075)$ | 1.79 | $(1.86,2.97)$ | $0.05(0.397)$ | 0.07 | $(-0.12,0.22)$ |

Table 23
2-PL MEX: Difficulty by AE Type

| AE | SRS of Size 30 |  |  | SRS of Size 250 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean | $\boldsymbol{N}$, Mean (SD) | Median | 95\% (2-sided) <br> CI for Mean |
| 1 | $1.12(0.571)$ | 1.16 | $(1.20,1.59)$ | $2.56(0.377)$ | 2.48 | $(2.41,2.71)$ |
| 2 | $1.17(0.541)$ | 1.17 | $(1.28,1.65)$ | $2.86(0.415)$ | 2.84 | $(2.70,3.03)$ |
| 3 | $1.19(0.579)$ | 1.21 | $(1.29,1.68)$ | $2.77(0.394)$ | 2.77 | $(2.61,2.93)$ |
| 4 | $1.02(0.574)$ | 1.07 | $(1.09,1.47)$ | $2.88(0.452)$ | 2.82 | $(2.70,3.06)$ |
| 5 | $1.20(0.559)$ | 1.21 | $(1.31,1.69)$ | $2.93(0.404)$ | 2.88 | $(2.77,3.09)$ |
| 6 | $1.20(0.551)$ | 1.22 | $(1.31,1.68)$ | $2.93(0.482)$ | 2.87 | $(2.73,3.12)$ |
| 7 | $1.14(0.541)$ | 1.09 | $(1.24,1.60)$ | $2.92(0.382)$ | 2.89 | $(2.76,3.07)$ |
| 8 | $1.17(0.585)$ | 1.14 | $(1.26,1.66)$ | $3.02(0.441)$ | 3.02 | $(2.84,3.20)$ |
| 9 | $1.12(0.543)$ | 1.10 | $(1.21,1.58)$ | $2.96(0.442)$ | 2.94 | $(2.78,3.14)$ |

acceptable patient discriminatory power because $\boldsymbol{A} \approx 0$ for all AE types (de Gruijter, 2004) and some IRFs were not monotonic. This finding was anticipated because the distribution of rare transfusion-related AEs was demonstrated to be non-normal in Figure 7 (p. 107). This finding means that statistical inference based on this model may not be reliable for the type of data under investigation.

For the 2-PL EX IRT model, the mean of the discrimination parameter ranged from 0.97 to 0.99 and 0.77 to 0.82 for $n=30$ and 250, respectively. The mean of the difficulty parameter ranged from 1.82 to 1.94 and 2.82 to 3.21 , respectively, for $n=30$ and 250. The mean of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from -0.63 to -0.38 and -0.16 to $\approx 0.00$ for $n=30$ and 250 , respectively. Trace and G-R plots for the three chains of starting values revealed consistent convergence of the mean and standard deviation of the discrimination, difficulty, and AE Predisposition ( $\boldsymbol{\theta}$ ) IRT parameters. Last, in conjunction with the posterior probability interval and residual plots, the 2-PL EX IRT model exhibited acceptable patient discriminatory power because $\boldsymbol{A}>0.75$ (de Gruijter, 2004) for all AE types. This result was anticipated because the distribution of rare transfusion-related AEs could be modeled with a gamma PDF (Figure 7, p. 107). This finding means that statistical inference based on this model was considered reliable for the type of data under investigation.

For the 2-PL MEX IRT model, the mean of the discrimination parameter ranged from 1.73 to 2.07 and -0.04 to 0.12 for $n=30$ and 250 , respectively. The mean difficulty parameter ranged from 1.02 to 1.20 and 2.56 to 3.02 for $n=30$ and 250 , respectively. The mean of the AE Predisposition $(\boldsymbol{\theta})$ parameter ranged from -0.79 to -0.50 and -0.45 to 0.40 for $n=30$ and 250 , respectively. Trace and G-R plots for the three chains of starting
values revealed inconsistent convergence of the mean and standard deviation of the discrimination, difficulty, and AE Predisposition ( $\boldsymbol{\theta}$ ) IRT parameters. Last, in conjunction with the posterior probability interval and residual plots, the 2-PL MEX IRT model exhibited acceptable patient discriminatory power for $n=30$ because $\boldsymbol{A}>0.75$. This result was anticipated because the distribution of rare transfusion-related AEs could also be modeled with a beta PDF (Figure 7, p. 107). Because GoF and parameter convergence was not consistent across all AE types and sample sizes, the statistical inference from this model for the type of data under investigation was considered unreliable.

In brief summary, 2-PL was not sufficient for modeling rare binomial events. 2-PL EX was sufficient for modeling this type of data. 2-PL MEX was not sufficient for modeling this type of data. As a result, statistical inference and power analysis results were based on the 2-PL EX IRT model.

Research Question 2 Results

This section investigates superiority and equivalence study objectives based on fixed historical rate data for the 2-PL EX IRT model. Analyses were based on simple random samples of sizes $n=30$ and 250. Paired Item Response Functions (IRFs) were first presented for each type of AE. Next, linear trapezoid and spline approximations were compared to the exact AUC for each type of AE. Statistical inference was then used to compare differences (Treatment A - Treatment B) between paired IRFs for superiority and equivalence study objectives. A parametric and nonparametric bootstrap, jackknife, and partial batch approach was used for estimation.

Each of these analyses was based on the mean and median of the paired effect and its standard error (SE). A 95\% confidence interval (CI) and its coverage were computed for each estimator. Next, the $t$-statistic and $p$-value were used to compare superiority hypotheses against zero and equivalence hypotheses against a $\Delta=10 \%$ margin. This margin was arbitrarily chosen, but Table 30 (p. 166) was constructed to assist researchers in deciding upon a margin for their study. Last, the Shapiro-Wilk's test was used to determine if the distribution of the paired differences in AUC were normal ( $p$-value $\geq 0.10$ ). The findings of this investigation were used to determine which set of statistics were reliable for performing statistical inference on superiority and equivalence study objectives. The statistical approach with the lowest bias, smallest standard error, and highest coverage was recommended for utilization for the analysis of the historical rate data, Data Patterns 1-4, and Research Question 3.

Figure 53 (p. 157) presents the Item Response Functions (IRFs) for each type of AE for $n=30$. This figure revealed that the empirical distribution of AE Predisposition $(\boldsymbol{\theta})$ for each AE type was comparable. As a result, the AE types were aggregated using the Partial approach (defined in Chapter II) so that analysis could be performed across all AE types. This approach was deemed reliable because the IRFs were comparable in distribution. If these IRFs possessed different distributions where different families of densities were present, this aggregation would require, for example, the use of transformation theory to construct a reliable composite for performing inference (Hogg, McKean, \& Craig, 2005).


Figure 53. IRT Plots for Fixed Historical Rate Data $(n=30)$

Table 24 (p. 158) presents the linear trapezoid and spline approximations to the exact AUC by treatment group and AE type for $n=30$ (Program 28.sas). For Treatment A, the exact area under IRFs ranged from 4.555 to 4.639 , and the median of the exact AUC across all AE types was 4.622. The linear approximation of the exact AUC ranged from 4.410 to 4.475, and the median (bias) of the approximation across all AE types was $4.622(-0.160)$. The spline approximation of the exact AUC ranged from 4.422 to 4.492, and the median (bias) of the approximation across all AE types was 4.477 ( -0.145 ). For Treatment B, the exact AUC ranged from 4.509 to 4.666 , and the median of the exact AUC across all AE types was 4.622. The linear approximation of the exact AUC ranged from 4.201 to 4.326, and the median (bias) of the approximation across all AE types was $4.290(-0.331)$. The spline approximation of the exact AUC ranged from 4.491 to 4.622,
and the median (bias) of the approximation across all AE types was $4.585(-0.036)$.
Additional analysis demonstrated that the Partial and Batch approaches for estimating AUC for were comparable for both AUC approximations. As a result, the Partial approach was utilized to aggregate the area under paired IRFs for all further analysis.

Table 24
Summary of Approximate Area Under IRFs by Treatment $(n=30)$

| AE | Treatment A |  |  | Treatment B |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Exact <br> AUC | Linear Trapezoid <br> AUC (Bias) | Spline AUC <br> (Bias) | Exact AUC | Linear Trapezoid <br> AUC (Bias) | Spline AUC <br> (Bias) |
| 1 | 4.55505 | $4.410(-0.145)$ | $4.422(-0.133)$ | 4.50948 | $4.201(-0.309)$ | $4.491(-0.018)$ |
| 2 | 4.60949 | $4.453(-0.157)$ | $4.467(-0.142)$ | 4.66623 | $4.326(-0.340)$ | $4.622(-0.045)$ |
| 3 | 4.62422 | $4.464(-0.160)$ | $4.480(-0.145)$ | 4.56258 | $4.243(-0.320)$ | $4.535(-0.027)$ |
| 4 | 4.59667 | $4.444(-0.153)$ | $4.458(-0.139)$ | 4.62224 | $4.290(-0.332)$ | $4.585(-0.037)$ |
| 5 | 4.63887 | $4.475(-0.163)$ | $4.492(-0.147)$ | 4.59470 | $4.269(-0.325)$ | $4.563(-0.032)$ |
| 6 | 4.60211 | $4.447(-0.155)$ | $4.461(-0.141)$ | 4.59255 | $4.267(-0.326)$ | $4.560(-0.033)$ |
| 7 | 4.62224 | $4.461(-0.161)$ | $4.477(-0.145)$ | 4.62422 | $4.293(-0.331)$ | $4.587(-0.037)$ |
| 8 | 4.63887 | $4.475(-0.163)$ | $4.492(-0.147)$ | 4.62601 | $4.295(-0.331)$ | $4.590(-0.036)$ |
| 9 | 4.62962 | $4.467(-0.163)$ | $4.483(-0.147)$ | 4.62962 | $4.296(-0.333)$ | $4.591(-0.039)$ |
| Xbar | 4.61302 | $4.455(-0.158)$ | $4.470(-0.143)$ | 4.60307 | $4.276(-0.327)$ | $4.569(-0.034)$ |
| Med | 4.62224 | $4.461(-0.160)$ | $4.477(-0.145)$ | 4.62224 | $4.290(-0.331)$ | $4.585(-0.036)$ |

Table 25 (p. 159) presents the linear trapezoid and spline approximations to the exact area under IRFs by treatment group and AE type for $n=250$ (Program 29.sas). For Treatment A, the exact AUC ranged from 5.072 to 5.211, and the median of the exact AUC across all AE types was 5.198. The linear approximation of the exact AUC ranged from 4.921 to 5.071 , and the median (bias) of the approximation across all AE types was
$5.058(-0.140)$. The spline approximation of the exact AUC ranged from 4.637 to 4.792, and the median (bias) of the approximation across all AE types was 4.777 (-0.420). For Treatment B, the exact AUC ranged from 5.060 to 5.222 , and the median of the exact AUC across all AE types was 5.191. The linear approximation of the exact AUC ranged from 4.783 to 4.970 , and the median (bias) of the approximation across all AE types was $4.943(-0.252)$. The spline approximation of the exact AUC ranged from 5.366 to 5.549, and the median (bias) of the approximation across all AE types was 5.522 (0.326).

## Table 25

Summary of Approximate Area Under IRFs by Treatment ( $n=250$ )

| AE | Treatment A |  |  |  | Treatment B |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Exact <br> AUC | Linear Trapezoid <br> AUC (Bias) | Spline AUC <br> (Bias) | Exact AUC | Linear Trapezoid <br> AUC (Bias) | Spline AUC <br> (Bias) |  |
| 1 | 5.07165 | $4.921(-0.151)$ | $4.637(-0.435)$ | 5.05977 | $4.783(-0.276)$ | $5.366(0.306)$ |  |
| 2 | 5.15794 | $5.013(-0.145)$ | $4.731(-0.427)$ | 5.12398 | $4.859(-0.265)$ | $5.439(0.315)$ |  |
| 3 | 5.16535 | $5.025(-0.140)$ | $4.744(-0.422)$ | 5.16842 | $4.918(-0.251)$ | $5.497(0.329)$ |  |
| 4 | 5.19779 | $5.058(-0.140)$ | $4.777(-0.420)$ | 5.17471 | $4.922(-0.253)$ | $5.501(0.326)$ |  |
| 5 | 5.18399 | $5.044(-0.140)$ | $4.763(-0.421)$ | 5.21305 | $4.960(-0.253)$ | $5.539(0.326)$ |  |
| 6 | 5.19779 | $5.058(-0.140)$ | $4.777(-0.420)$ | 5.19594 | $4.948(-0.248)$ | $5.527(0.331)$ |  |
| 7 | 5.20689 | $5.067(-0.140)$ | $4.787(-0.420)$ | 5.21591 | $4.967(-0.249)$ | $5.545(0.329)$ |  |
| 8 | 5.20689 | $5.067(-0.140)$ | $4.787(-0.420)$ | 5.19140 | $4.943(-0.249)$ | $5.522(0.330)$ |  |
| 9 | 15.21141 | $5.071(-0.140)$ | $4.792(-0.420)$ | 5.22211 | $4.970(-0.252)$ | $5.549(0.326)$ |  |
| Xbar | 5.17774 | $5.036(-0.142)$ | $4.755(-0.423)$ | 5.17392 | $4.919(-0.255)$ | $5.498(0.324)$ |  |
| Med | 5.19779 | $5.058(-0.140)$ | $4.777(-0.420)$ | 5.19140 | $4.943(-0.252)$ | $5.522(0.326)$ |  |

Using the overall median bias ("Med") from Table 24 (p. 158) and Table 25
(p. 159) across treatment groups, the median of these biases was -0.206 for the linear
trapezoid approximation and -0.09 for the spline approximation. As a result, the spline approximation better approximated the exact AUC for the investigated AE data.

## Superiority Analysis

Tables 26-27 (p. 161) present the superiority analyses for comparing the AE Predisposition ( $\boldsymbol{\theta}$ ) IRFs in Figure 53 (p. 157). Before the hypothesis test results are discussed for $n=30$, the coverage (\%) or observed confidence level was inspected for each type of estimation. For the linear trapezoid approximation to the exact AUC, the coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $81.50 \%, 77.25 \%$, and $61.20 \%$, respectively. The coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $95.57 \%, 50.00 \%$, and $97.48 \%$, respectively. For the spline approximation to the exact AUC, the coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $89.05 \%, 99.40 \%$, and $92.01 \%$, respectively. The coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $95.51 \%, 50.00 \%$, and $97.48 \%$, respectively.

These results demonstrated that coverage achieved $95 \%$ for the bootstrap and partial batch approaches based on the median. For these approaches, coverage was larger for the median because the distribution of the effects was consistently non-normal. As a result, the bootstrap and partial batch approaches based on the median will be highlighted in the next section for the spline approximation to the exact AUC. The above analyses for each AE type are repeated in Appendix K.

For the bootstrap approach, the median (SE) of the paired difference or effect was $1.088 \times 10^{-13}\left(2.163 \times 10^{-13}\right)$. The $95 \%$ (two-sided) CI on this effect was $\left(-9.864 \times 10^{-14}\right.$,

Table 26
Superiority Analysis: Linear Trapezoid Approximation ( $n=30$ )

| Type of Estimation | Estimator | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { CI on } \\ \text { Difference } \end{gathered}$ | Coverage (\%) | Tstat ( $p$-value) | S-W ${ }^{1}$ Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | $\begin{gathered} 1.371 \mathrm{E}-04 \\ (5.356 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.077 \mathrm{E}-02 \\ 1.105 \mathrm{E}-02) \end{gathered}$ | 81.50 | $\begin{gathered} 8.440 \mathrm{E}-02 \\ (0.370) \end{gathered}$ | $\begin{gathered} 0.5656 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.101 \mathrm{E}-13 \\ (2.285 \mathrm{E}-13) \end{gathered}$ | $\begin{gathered} (-9.908 \mathrm{E}-14 \\ 2.643 \mathrm{E}-13) \end{gathered}$ | 95.57 | $\begin{gathered} 1.963 \mathrm{E}+00 \\ (0.089) \end{gathered}$ |  |
| Jackknife ${ }^{3}$ | Mean | $\begin{gathered} 1.471 \mathrm{E}-04 \\ (6.743 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-1.226 \mathrm{E}-03 \\ 1.521 \mathrm{E}-03) \end{gathered}$ | 77.25 | $\begin{gathered} 1.950 \mathrm{E}-01 \\ (0.455) \end{gathered}$ | $\begin{gathered} 0.5680 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.097 \mathrm{E}-13 \\ (6.098 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-1.693 \mathrm{E}-03, \\ 1.693 \mathrm{E}-03) \end{gathered}$ | 50.00 | $\begin{gathered} 5.522 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| Partial <br> Batch ${ }^{4}$ | Mean | $\begin{gathered} 1.471 \mathrm{E}-04 \\ (1.990 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-3.906 \mathrm{E}-03 \\ 4.201 \mathrm{E}-03) \end{gathered}$ | 61.20 | $\begin{gathered} 1.805 \mathrm{E}-02 \\ (0.776) \end{gathered}$ | $\begin{gathered} 0.5700 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.101 \mathrm{E}-13 \\ (5.773 \mathrm{E}-14) \end{gathered}$ | $\begin{gathered} (-1.415 \mathrm{E}-13 \\ 1.576 \mathrm{E}-13) \end{gathered}$ | 97.48 | $\begin{gathered} 1.198 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas; ${ }^{4}$ Program 32.sas

## Table 27

Superiority Analysis: Spline Approximation $(n=30)$

| Type of Estimation | Estimator | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { CI on } \\ \text { Difference } \end{gathered}$ | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p \text {-value) } \end{gathered}$ | S-W ${ }^{1}$ Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | $\begin{aligned} & -8.446 \mathrm{E}-03 \\ & (6.787 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-2.227 \mathrm{E}-02 \\ 5.379 \mathrm{E}-03) \end{gathered}$ | 89.05 | $\begin{gathered} -3.236 \mathrm{E}+00 \\ (0.219) \end{gathered}$ | $\begin{gathered} 0.4350 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.088 \mathrm{E}-13 \\ (2.163 \mathrm{E}-13) \end{gathered}$ | $\begin{gathered} (-9.864 \mathrm{E}-14 \\ 2.100 \mathrm{E}-13) \end{gathered}$ | 95.51 | $\begin{gathered} 1.958 \mathrm{E}+00 \\ (0.090) \end{gathered}$ |  |
| Jackknife ${ }^{3}$ | Mean | $\begin{aligned} & -8.509 \mathrm{E}-03 \\ & (3.083 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-1.479 \mathrm{E}-02, \\ -2.229 \mathrm{E}-03) \end{gathered}$ | 99.40 | $\begin{gathered} -2.764 \mathrm{E}+00 \\ (0.012) \end{gathered}$ | $\begin{gathered} 0.4165 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 9.348 \mathrm{E}-14 \\ (8.372 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-2.324 \mathrm{E}-02, \\ 2.324 \mathrm{E}-02) \end{gathered}$ | 50.00 | $\begin{gathered} 1.074 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| Partial <br> Batch ${ }^{4}$ | Mean | $\begin{aligned} & -8.509 \mathrm{E}-03 \\ & (5.816 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-2.036 \mathrm{E}-02, \\ 3.337 \mathrm{E}-03) \end{gathered}$ | 92.01 | $\begin{gathered} -1.492 \mathrm{E}+00 \\ (0.160) \end{gathered}$ | $\begin{gathered} 0.4154 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.088 \mathrm{E}-13 \\ (5.795 \mathrm{E}-14) \end{gathered}$ | $\begin{gathered} (-1.421 \mathrm{E}-13, \\ 2.087 \mathrm{E}-13) \end{gathered}$ | 97.48 | $\begin{gathered} 1.198 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |

[^0]$2.100 \times 10^{-13}$ ). The test statistic ( $p$-value) was $1.958(0.090)$ for a one-sample $t$-distribution. This result demonstrated that the null hypothesis of no paired difference between treatment groups was not rejected (i.e., the CI contained zero). That is, Treatment A was not statistically different from or superior to Treatment B at $95 \%$ confidence. These results were consistent with partial batch estimation.

Tables 28-29 (p. 163) present the superiority analyses for comparing the AE Predisposition ( $\boldsymbol{\theta}$ ) IRFs in Figure 54 (p. 164). Before the hypothesis test results are discussed for $n=250$, the coverage (\%) was inspected for each type of estimation. For the linear trapezoid approximation to the exact AUC, the coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $62.58 \%, 88.48 \%$, and $52.25 \%$, respectively. The coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $>99.99 \%, 50.00 \%$, and $>99.99 \%$, respectively. For the spline approximation to the exact AUC, the coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $75.97 \%,>99.99 \%$, and $59.91 \%$, respectively. The coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $>99.99 \%, 50.00 \%$, and $>99.99 \%$, respectively.

These results demonstrated that coverage achieved $95 \%$ for the bootstrap and partial batch approaches based on the median. For these approaches, coverage was larger for the median because the distribution of the effects was consistently non-normal. As a result, the bootstrap and partial batch approaches based on the median will be discussed in the next section for the spline approximation to the exact AUC. The above analyses for each AE type are repeated in Appendix K.

Table 28
Superiority Analysis: Linear Trapezoid Approximation ( $n=250$ )

| Type of Estimation | Estimator | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { CI on } \\ \text { Difference } \end{gathered}$ | Coverage (\%) | Tstat (p-value) | S-W ${ }^{1}$ Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | $\begin{aligned} & -1.089 \mathrm{E}-04 \\ & (3.073 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-6.161 \mathrm{E}-03 \\ 5.943 \mathrm{E}-03) \end{gathered}$ | 62.58 | $\begin{gathered} 1.986 \mathrm{E}-02 \\ (0.748) \end{gathered}$ | $\begin{gathered} 0.2753 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} -4.130 \mathrm{E}-14 \\ (2.665 \mathrm{E}-14) \end{gathered}$ | $\begin{gathered} (-9.893 \mathrm{E}-14 \\ 1.544 \mathrm{E}-14) \end{gathered}$ | >99.99 | $\begin{gathered} -1.563 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| Jackknife ${ }^{3}$ | Mean | $\begin{aligned} & -1.154 \mathrm{E}-04 \\ & (6.568 \mathrm{E}-05) \end{aligned}$ | $\begin{gathered} (-2.448 \mathrm{E}-04 \\ 1.396 \mathrm{E}-05) \end{gathered}$ | 88.48 | $\begin{gathered} -1.810 \mathrm{E}+00 \\ (0.230) \end{gathered}$ | $\begin{gathered} 0.2029 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -4.130 \mathrm{E}-14 \\ & (1.279 \mathrm{E}-04) \end{aligned}$ | $\begin{gathered} (-2.743 \mathrm{E}-04, \\ 2.743 \mathrm{E}-04) \end{gathered}$ | 50.00 | $\begin{gathered} -2.784 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| Partial <br> Batch ${ }^{4}$ | Mean | $\begin{aligned} & -1.154 \mathrm{E}-04 \\ & (2.088 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-4.227 \mathrm{E}-03, \\ 3.996 \mathrm{E}-03) \end{gathered}$ | 52.25 | $\begin{gathered} -5.618 \mathrm{E}-02 \\ (0.955) \end{gathered}$ | $\begin{gathered} 0.2024 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -4.130 \mathrm{E}-14 \\ & (2.465 \mathrm{E}-14) \end{aligned}$ | $\begin{aligned} & (-4.273 \mathrm{E}-14, \\ & -2.877 \mathrm{E}-14) \end{aligned}$ | >99.99 | $\begin{gathered} -3.064 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |

${ }^{1}$ Shapiro-Wilk test for normality of paired AUC differences; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas;
${ }^{4}$ Program 32.sas

Table 29
Superiority Analysis: Spline Approximation ( $n=250$ )

| Type of Estimation | Estimator | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { CI on } \\ \text { Difference } \end{gathered}$ | Coverage (\%) | Tstat (p-value) | $\begin{gathered} \text { S-W }{ }^{1} \\ \begin{array}{c} \text { Stat } \\ (p \text {-value }) \end{array} \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | $\begin{aligned} & -3.498 \mathrm{E}-03 \\ & (1.133 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-2.580 \mathrm{E}-02 \\ 1.881 \mathrm{E}-02) \end{gathered}$ | 75.97 | $\begin{gathered} -6.786 \mathrm{E}-01 \\ (0.481) \end{gathered}$ | $\begin{gathered} 0.1491 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -4.174 \mathrm{E}-14 \\ & (2.620 \mathrm{E}-14) \end{aligned}$ | $\begin{gathered} (-9.913 \mathrm{E}-14 \\ 1.540 \mathrm{E}-14) \end{gathered}$ | >99.99 | $\begin{gathered} -1.567 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| Jackknife ${ }^{3}$ | Mean | $\begin{aligned} & -3.504 \mathrm{E}-03 \\ & (4.722 \mathrm{E}-04) \end{aligned}$ | $\begin{aligned} & (-4.433 \mathrm{E}-03, \\ & -2.574 \mathrm{E}-03) \end{aligned}$ | >99.99 | $\begin{gathered} -7.421 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0879 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -4.174 \mathrm{E}-14 \\ & (3.506 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-7.519 \mathrm{E}-03, \\ 7.519 \mathrm{E}-03) \end{gathered}$ | 50.00 | $\begin{gathered} -1.182 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| Partial <br> Batch ${ }^{4}$ | Mean | $\begin{aligned} & -3.504 \mathrm{E}-03 \\ & (1.394 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-3.096 \mathrm{E}-02 \\ 2.395 \mathrm{E}-02) \end{gathered}$ | 59.91 | $\begin{gathered} -2.514 \mathrm{E}-01 \\ (0.802) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -4.174 \mathrm{E}-14 \\ & (2.620 \mathrm{E}-14) \end{aligned}$ | $\begin{gathered} (-9.889 \mathrm{E}-14 \\ 1.540 \mathrm{E}-14) \end{gathered}$ | >99.99 | $\begin{gathered} -1.567 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |

1 Shapiro-Wilk test for normality of the effect; 2 Program 30.sas; 3 Program 31.sas; 4 Program 32.sas


Figure 54. IRT Plots for Fixed Historical Rate Data ( $n=250$ )

For the bootstrap approach, the median (SE) of the paired difference was $-4.174 \times 10^{-14}\left(2.620 \times 10^{-14}\right)$. The $95 \%$ (two-sided) CI on this effect was $\left(-9.913 \times 10^{-14}\right.$, $1.540 \times 10^{-14}$ ). The test statistic ( $p$-value) was $-1.567(<0.001)$ for a one-sample $t$-distribution. The conclusion based on this $p$-value was inconsistent between the CI and associated $p$-value. The $p$-value demonstrated that Treatment A was statistically superior (i.e., negative difference) to Treatment B at $95 \%$ confidence. The opposite and correct conclusion was reached with the CI because it contained zero. This inconsistency was likely due to an artificial effect caused by sample size. Statistical substantiation of superiority was not warranted because of the IRF overlap in Figure 54. These results and conclusions were consistent with partial batch estimation. The above inconsistency will be further discussed in Chapter V.

## Equivalence Margin Evaluation

Opposite to superiority which is based on the minimum acceptable difference between treatment groups, equivalence is based on the maximum acceptable difference. This difference is commonly denoted by the equivalence margin $\Delta$. Seven definitions for computing this margin were presented in this study. Unfortunately, there is no consensus on which definition to use. As a result, this study performed simulations to assist with this choice for treatment differences in binomial event data. Definitions that result in margins that exceed the numerical difference between treatments and decrease as a function of this definition are recommended for further consideration.

Monte Carlo simulations based on 5,000 nonparametric bootstraps were performed to investigate equivalence margins for differences in AEs that ranged from $0.00039 \%$ to $2.609 \%$ (Program 33.sas).

Table 30 (p. 166) presents the simulated equivalence margins for the 7 definitions. All values for Margin 1 fell below the treatment group difference. Margin 2 failed to exceed this difference for rates larger than 0.05. Margin 4 (linear trapezoid and spline) failed to exceed this difference for rates smaller than 0.0012. Margin 6 (FDA definition) exceeded this difference for all rates investigated, but these margins did not decrease with the magnitude of the percent difference between treatments A and B. Margins 2 and 5 exceeded this difference for all rates investigated and these margins decreased with the magnitude of the percent difference between treatments A and B. As a result, Margin 2 and Margin 5 were the only viable equivalence margin options for the range of data
investigated. Furthermore, there is no existing criteria for being able to choose Margin 2 over Margin 5, and vice versa.

Table 30

## Evaluation of Various Equivalence Margins

| Test Group (\%) | $\begin{aligned} & \text { Control } \\ & \text { Group } \\ & (\%) \end{aligned}$ | Diff (\%) | Margin (\%) |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | 1 | 2 | 3 | Linear Trapezoid | Spline | 5 | 6 |
| 20 | 17.391 | 2.609 | 0.5 | 7.8 | 1.3 | 16.7 | 16.7 | 8.9 | 4.3 |
| 15 | 13.0433 | 1.95675 | 0.4 | 6.9 | 1.1 | 12.5 | 12.5 | 7.9 | 3.6 |
| 10 | 8.6955 | 1.30450 | 0.3 | 5.8 | 0.9 | 9.1 | 9.1 | 6.6 | 3.7 |
| 5 | 4.3478 | 0.65225 | 0.1 | 4.2 | 0.6 | 4.8 | 4.8 | 4.8 | 4.0 |
| 1 | 0.8696 | 0.13045 | 0.03 | 1.9 | 0.3 | 0.9 | 0.95 | 2.2 | 4.8 |
| 0.66 | 0.5739 | 0.08610 | 0.02 | 1.6 | 0.2 | 0.5 | 0.58 | 1.8 | 4.4 |
| 0.266 | 0.2313 | 0.03470 | 0.01 | 1.0 | 0.1 | 0.1 | 0.10 | 1.1 | 3.8 |
| 0.120 | 0.1043 | 0.01565 | 0.003 | 0.7 | 0.09 | 0.002 | 0.002 | 0.8 | 4.0 |
| 0.053 | 0.0461 | 0.00691 | 0.001 | 0.5 | 0.06 | $1.106 \times 10^{-7}$ | $1.254 \times 10^{-7}$ | 0.5 | 1.0 |
| 0.0489 | 0.0425 | 0.00638 | 0.001 | 0.4 | 0.06 | $2.449 \times 10^{-8}$ | $2.777 \times 10^{-8}$ | 0.5 | 1.3 |
| 0.0168 | 0.0146 | 0.00219 | 0.0004 | 0.3 | 0.04 | $6.243 \times 10^{-25}$ | $7.080 \times 10^{-25}$ | 0.3 | 0.6 |
| 0.0057 | 0.0049 | 0.00074 | 0.0001 | 0.2 | 0.02 | $9.892 \times 10^{-76}$ | $1.122 \times 10^{-75}$ | 0.2 | 4.2 |
| 0.0040 | 0.0035 | 0.00052 | 0.0001 | 0.1 | 0.02 | $1.915 \times 10^{-107}$ | $2.172 \times 10^{-107}$ | 0.1 | 1.2 |
| 0.0030 | 0.0026 | 0.00039 | 0.0001 | 0.1 | 0.02 | $1.074 \times 10^{-143}$ | $1.217 \times 10^{-143}$ | 0.1 | 1.3 |

When this situation arises, the next step is to typically compare sample size requirements between these margins. If the sample size estimates for both margins are financially feasible, the more conservative margin may be used. If the sample size based on one margin is too large, the other margin should be used. After a margin is identified for use, it should be determined if it is clinically or scientifically reasonable with respect to the impact of the study endpoint. This subjective determination is beyond the scope of
this study. For this study, all equivalence analyses were based on an arbitrary equivalence margin of $10 \%$.

## Equivalence Analysis

Table 31 (p. 168) and Table 32 (p. 169) present the equivalence analyses for comparing the AE Predisposition ( $\boldsymbol{\theta}$ ) IRFs in Figure 53 (p. 157). Before the hypothesis test results are discussed for $n=30$, the coverage (\%) or observed confidence level was inspected for each type of estimation. For the linear trapezoid approximation to the exact AUC, the minimum coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $91.74 \%,>99.99 \%$, and $99.90 \%$, respectively. The minimum coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $94.35 \%, 50.00 \%$, and $94.26 \%$, respectively. For the spline approximation to the exact AUC, the minimum coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $80.08 \%, 91.15 \%$, and $73.30 \%$, respectively. The minimum coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $96.41 \%, 50.00 \%$, and $94.34 \%$, respectively.

These results demonstrated that minimum coverage achieved $95 \%$ for the jackknife and partial batch approaches based on the mean and the bootstrap approach based on the median. The distribution of the lower and upper effects was consistently non-normal. As a result, the bootstrap approach based on the median will be highlighted in the next section for the spline approximation to the exact area under IRFs. The above analyses for each AE type are repeated in Appendix K.

Table 31
Equivalence Analysis: Linear Trapezoid Approximation ( $n=30$ )

| Type of Estimation | Estimator | Effect | $\begin{aligned} & \text { Difference } \\ & (S E) \end{aligned}$ | $\begin{aligned} & \text { 95\% (2-sided) } \\ & \text { CI on } \\ & \text { Difference } \end{aligned}$ | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p \text {-value }) \end{gathered}$ | $\underset{(p-\text {-value })}{\text { S-W }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | Lower | $\begin{aligned} & -1.335 \mathrm{E}-02 \\ & (6.017 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-2.561 \mathrm{E}-02, \\ & -1.095 \mathrm{E}-03) \end{aligned}$ | 92.66 | $\begin{gathered} -5.973 \mathrm{E}+00 \\ (0.147) \end{gathered}$ | $\begin{gathered} 0.5053 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 1.365 \mathrm{E}-02 \\ (6.680 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (4.530 \mathrm{E}-05, \\ 2.726 \mathrm{E}-02) \end{gathered}$ | 91.74 | $\begin{gathered} 3.849 \mathrm{E}+00 \\ (0.165) \end{gathered}$ | $\begin{gathered} 0.4818 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.168 \mathrm{E}-12 \\ & (4.511 \mathrm{E}-12) \end{aligned}$ | $\begin{gathered} (-2.169 \mathrm{E}-11, \\ 3.285 \mathrm{E}-12) \end{gathered}$ | 94.49 | $\begin{gathered} -2.046 \mathrm{E}+00 \\ (0.110) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.329 \mathrm{E}-12 \\ (4.640 \mathrm{E}-12) \end{gathered}$ | $\begin{gathered} (-3.554 \mathrm{E}-12, \\ 3.125 \mathrm{E}-11) \end{gathered}$ | 94.35 | $\begin{gathered} 2.023 \mathrm{E}+00 \\ (0.113) \end{gathered}$ | N/A |
| Jackknife ${ }^{3}$ | Mean | Lower | $\begin{aligned} & -1.334 \mathrm{E}-02 \\ & (1.668 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-1.674 \mathrm{E}-02, \\ -9.941 \mathrm{E}-03) \end{gathered}$ | >99.99 | $\begin{gathered} -7.989 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5100 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 1.363 \mathrm{E}-02 \\ (1.670 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (1.023 \mathrm{E}-02, \\ 1.703 \mathrm{E}-02) \end{gathered}$ | >99.99 | $\begin{gathered} 8.205 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.4848 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.164 \mathrm{E}-12 \\ & (1.317 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-3.657 \mathrm{E}-02, \\ 3.657 \mathrm{E}-02) \end{gathered}$ | 50.00 | $\begin{gathered} -7.037 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 1.389 \mathrm{E}-11 \\ (1.342 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-3.725 \mathrm{E}-02, \\ 3.725 \mathrm{E}-02) \end{gathered}$ | 50.00 | $\begin{gathered} 1.020 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ | N/A |
| Partial Batch ${ }^{4}$ | Mean | Lower | $\begin{aligned} & -1.334 \mathrm{E}-02 \\ & (2.593 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-1.862 \mathrm{E}-02, \\ -8.056 \mathrm{E}-03) \end{gathered}$ | 99.98 | $\begin{gathered} -5.764 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5102 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 1.363 \mathrm{E}-02 \\ (3.771 \mathrm{E}-03) \end{gathered}$ | $\begin{aligned} & (5.952 \mathrm{E}-03, \\ & 2.131 \mathrm{E}-02) \end{aligned}$ | 99.90 | $\begin{gathered} 3.820 \mathrm{E}+00 \\ (0.002) \end{gathered}$ | $\begin{gathered} 0.4851 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.168 \mathrm{E}-12 \\ & (4.511 \mathrm{E}-12) \end{aligned}$ | $\begin{gathered} (-2.169 \mathrm{E}-11, \\ 3.356 \mathrm{E}-12) \end{gathered}$ | 94.41 | $\begin{gathered} -2.033 \mathrm{E}+00 \\ (0.112) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.329 \mathrm{E}-12 \\ (4.640 \mathrm{E}-12) \end{gathered}$ | $\begin{gathered} (-3.554 \mathrm{E}-12, \\ 2.221 \mathrm{E}-11) \end{gathered}$ | 94.26 | $\begin{gathered} 2.011 \mathrm{E}+00 \\ (0.115) \end{gathered}$ | N/A |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas; ${ }^{4}$ Program 32.sas

For the bootstrap approach in Table 32 (p. 169), the median (SE) of the lower effect was $-9.162 \times 10^{-12}\left(3.654 \times 10^{-12}\right)$. The $97.5 \%$ (one-sided) UCL on this effect was $3.437 \times 10^{-13}$. The test statistic ( $p$-value) on this effect was -2.681 ( 0.055 ) for a onesample $t$-distribution. The median (SE) of the upper effect was $9.294 \times 10^{-12}\left(4.597 \times 10^{-12}\right)$. The $97.5 \%$ (one-sided) LCL on this effect was $-3.435 \times 10^{-12}$. The test statistic ( $p$-value) on this effect was $2.433(0.072)$ for a one-sample $t$-distribution. In agreement with the
confidence limits, these $p$-values demonstrated that Treatment A was not statistically equivalent to Treatment B at $\Delta=10 \%$ for $95 \%$ confidence (two-sided). This lack of significance was likely due to insufficient statistical power. This means that the desired effect was numerically demonstrated to be within $\Delta=10 \%$, but the sample size caused the confidence limits to exceed this margin.

Table 32
Equivalence Analysis: Spline Approximation ( $n=30$ )

| Type of Estimation | Estimator | Effect | Difference (SE) | $\begin{aligned} & \text { 95\% (2-sided) } \\ & \text { CI on Difference } \end{aligned}$ | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p \text {-value }) \end{gathered}$ | $\underset{(p \text {-value })}{S-W^{1} \text { Stat }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | Lower | $\begin{aligned} & -2.290 \mathrm{E}-02 \\ & (1.091 \mathrm{E}-02) \end{aligned}$ | $\begin{aligned} & (-4.513 \mathrm{E}-02, \\ & -6.668 \mathrm{E}-04) \end{aligned}$ | 86.48 | $\begin{gathered} -2.875 \mathrm{E}+00 \\ (0.270) \end{gathered}$ | $\begin{gathered} 0.4050 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 5.947 \mathrm{E}-03 \\ (1.182 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.814 \mathrm{E}-02, \\ 3.003 \mathrm{E}-02) \end{gathered}$ | 80.08 | $\begin{gathered} 2.214 \mathrm{E}+00 \\ (0.398) \end{gathered}$ | $\begin{gathered} 0.4479 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.162 \mathrm{E}-12 \\ & (3.654 \mathrm{E}-12) \end{aligned}$ | $\begin{gathered} (-1.899 \mathrm{E}-11, \\ 3.437 \mathrm{E}-13) \end{gathered}$ | 97.24 | $\begin{gathered} -2.681 \mathrm{E}+00 \\ (0.055) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.294 \mathrm{E}-12 \\ (4.597 \mathrm{E}-12) \end{gathered}$ | $\begin{gathered} (-3.435 \mathrm{E}-12, \\ 2.213 \mathrm{E}-11) \end{gathered}$ | 96.41 | $\begin{gathered} 2.433 \mathrm{E}+00 \\ (0.072) \end{gathered}$ | N/A |
| Jackknife $^{3}$ | Mean | Lower | $\begin{aligned} & -2.291 \mathrm{E}-02 \\ & (4.187 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.143 \mathrm{E}-02, \\ & -1.438 \mathrm{E}-02) \end{aligned}$ | >99.99 | $\begin{gathered} -5.471 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.2931 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 5.887 \mathrm{E}-03 \\ (4.196 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-2.659 \mathrm{E}-03, \\ 1.443 \mathrm{E}-02) \end{gathered}$ | 91.15 | $\begin{gathered} 1.404 \mathrm{E}+00 \\ (0.177) \end{gathered}$ | $\begin{gathered} 0.4437 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.139 \mathrm{E}-12 \\ & (2.258 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-6.268 \mathrm{E}-02, \\ 6.268 \mathrm{E}-02) \end{gathered}$ | 50.00 | $\begin{gathered} -4.054 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.311 \mathrm{E}-12 \\ (5.778 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.604 \mathrm{E}-02, \\ 1.604 \mathrm{E}-02) \end{gathered}$ | 50.00 | $\begin{gathered} 1.601 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ | N/A |
| Partial Batch ${ }^{4}$ | Mean | Lower | $\begin{aligned} & -2.291 \mathrm{E}-02 \\ & (6.582 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-3.631 \mathrm{E}-02, \\ -9.499 \mathrm{E}-03) \end{gathered}$ | 99.91 | $\begin{gathered} -3.487 \mathrm{E}+00 \\ (0.002) \end{gathered}$ | $\begin{gathered} 0.2815 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 5.887 \mathrm{E}-03 \\ (9.494 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.345 \mathrm{E}-02, \\ 2.523 \mathrm{E}-02) \end{gathered}$ | 73.30 | $\begin{gathered} 6.359 \mathrm{E}-01 \\ (0.534) \end{gathered}$ | $\begin{gathered} 0.4441 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.162 \mathrm{E}-12 \\ & (4.818 \mathrm{E}-14) \end{aligned}$ | $\begin{array}{r} (-9.293 \mathrm{E}-12, \\ -9.030 \mathrm{E}-12) \end{array}$ | >99.99 | $\begin{gathered} -1.906 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.294 \mathrm{E}-12 \\ (4.597 \mathrm{E}-12) \end{gathered}$ | $\begin{gathered} (-3.469 \mathrm{E}-12, \\ 2.206 \mathrm{E}-11) \end{gathered}$ | 94.34 | $\begin{gathered} 2.023 \mathrm{E}+00 \\ (0.113) \end{gathered}$ | N/A |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas; ${ }^{4}$ Program 32.sas

Table 33 (p. 171) and Table 34 (p. 172) present the equivalence analyses for comparing the AE Predisposition ( $\boldsymbol{\theta}$ ) IRFs in Figure 54 (p. 164). Before the hypothesis test results are discussed for $n=250$, the coverage (\%) was inspected for each type of estimation. For the linear trapezoid approximation to the exact AUC, the minimum coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $74.43 \%,>99.99 \%$, and $77.47 \%$, respectively. The minimum coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $>99.99 \%, 50.00 \%$, and $>99.99 \%$, respectively. For the spline approximation to the exact AUC, the minimum coverage for the bootstrap, jackknife, and partial batch mean estimation approaches was $72.62 \%, 99.93 \%$, and $53.39 \%$, respectively. The minimum coverage for the bootstrap, jackknife, and partial batch median estimation approaches was $>99.99 \%, 50.00 \%$, and $>99.99 \%$, respectively. These results demonstrated that minimum coverage achieved $95 \%$ for the jackknife approach for the mean and the bootstrap and partial batch approaches for the median. The distribution of the lower and upper effects was consistently non-normal. As a result, the bootstrap approach based on the median will be highlighted in the next section for the spline approximation to the exact AUC. The above analyses for each AE type are repeated in Appendix K.

For the bootstrap approach, the median $(S E)$ of the lower paired effect was $-9.273 \times 10^{-12}\left(2.882 \times 10^{-14}\right)$. The $97.5 \%$ (one-sided) UCL on this effect was $-9.195 \times 10^{-12}$. The test statistic ( $p$-value) on this effect was $-3.130 \times 10^{2}(<0.001)$ for the one-sample $t$-distribution. The median $(S E)$ of the upper paired effect was $9.189 \times 10^{-12}\left(2.096 \times 10^{-14}\right)$. The $97.5 \%$ (one-sided) LCL on this effect was $9.150 \times 10^{-12}$. The test statistic ( $p$-value) on this effect was $4.444 \times 10^{2}(<0.001)$ for the one-sample $t$-distribution. In agreement with
the confidence limits, these $p$-values demonstrated that Treatment A was statistically equivalent to Treatment B for a 10\% equivalence margin with $95 \%$ confidence (twosided). That is, the upper confidence limit was less than zero for the lower paired effect and the lower confidence limit was greater than zero for the upper paired effect.

## Table 33

Equivalence Analysis: Linear Trapezoid Approximation $(n=250)$

| Type of Estimation | Estimator | Effect | Difference (SE) | 95\% (2-sided) CI on Difference | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p \text {-value }) \end{gathered}$ | S-W ${ }^{1}$ Stat ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | Lower | $\begin{aligned} & -2.117 \mathrm{E}-03 \\ & (2.220 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-6.490 \mathrm{E}-03, \\ 2.256 \mathrm{E}-03) \end{gathered}$ | 90.12 | $\begin{gathered} -1.992 \mathrm{E}+00 \\ (0.198) \end{gathered}$ | $\begin{gathered} 0.3575 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 1.887 \mathrm{E}-03 \\ (3.363 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-4.737 \mathrm{E}-03, \\ 8.511 \mathrm{E}-03) \end{gathered}$ | 74.43 | $\begin{gathered} 8.754 \mathrm{E}-01 \\ (0.511) \end{gathered}$ | $\begin{gathered} 0.1783 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.272 \mathrm{E}-12 \\ & (2.882 \mathrm{E}-14) \end{aligned}$ | $\begin{array}{r} (-9.336 \mathrm{E}-12, \\ -9.195 \mathrm{E}-12) \end{array}$ | >99.99 | $\begin{gathered} -3.103 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.190 \mathrm{E}-12 \\ (2.123 \mathrm{E}-14) \end{gathered}$ | $\begin{aligned} & (9.179 \mathrm{E}-12, \\ & 9.239 \mathrm{E}-12) \end{aligned}$ | >99.99 | $\begin{gathered} 4.428 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
| Jackknife ${ }^{3}$ | Mean | Lower | $\begin{aligned} & -2.117 \mathrm{E}-03 \\ & (4.828 \mathrm{E}-05) \end{aligned}$ | $\begin{aligned} & (-2.212 \mathrm{E}-03, \\ & -2.022 \mathrm{E}-03) \end{aligned}$ | >99.99 | $\begin{gathered} -4.388 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.2791 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 1.887 \mathrm{E}-03 \\ (4.953 \mathrm{E}-05) \end{gathered}$ | $\begin{gathered} (1.789 \mathrm{E}-03, \\ 1.984 \mathrm{E}-03) \end{gathered}$ | >99.99 | $\begin{gathered} 3.810 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.1177 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.272 \mathrm{E}-12 \\ & (2.136 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-4.581 \mathrm{E}-03, \\ 4.581 \mathrm{E}-03) \end{gathered}$ | 50.00 | $\begin{gathered} -4.374 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.208 \mathrm{E}-12 \\ (1.857 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-3.983 \mathrm{E}-03, \\ 3.983 \mathrm{E}-03) \end{gathered}$ | 50.00 | $\begin{gathered} 4.904 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ | N/A |
| Partial Batch ${ }^{4}$ | Mean | Lower | $\begin{aligned} & -2.117 \mathrm{E}-03 \\ & (1.104 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} \hline(-4.291 \mathrm{E}-03, \\ 5.623 \mathrm{E}-05) \end{gathered}$ | 97.02 | $\begin{gathered} -1.938 \mathrm{E}+00 \\ (0.060) \end{gathered}$ | $\begin{gathered} 0.2783 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{gathered} 1.887 \mathrm{E}-03 \\ (2.499 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-3.035 \mathrm{E}-03, \\ 6.809 \mathrm{E}-03) \end{gathered}$ | 77.47 | $\begin{gathered} 7.561 \mathrm{E}-01 \\ (0.451) \end{gathered}$ | $\begin{gathered} 0.1177 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.272 \mathrm{E}-12 \\ & (3.191 \mathrm{E}-14) \end{aligned}$ | $\begin{aligned} & (-9.341 \mathrm{E}-12, \\ & -9.195 \mathrm{E}-12) \end{aligned}$ | >99.99 | $\begin{gathered} -2.892 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.190 \mathrm{E}-12 \\ (2.123 \mathrm{E}-14) \end{gathered}$ | $\begin{aligned} & (9.144 \mathrm{E}-12, \\ & 9.235 \mathrm{E}-12) \end{aligned}$ | >99.99 | $\begin{gathered} 4.410 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |

[^1]Table 34
Equivalence Analysis: Spline Approximation ( $n=250$ )

| Type of Estimation | Estimator | Effect | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { ander } \end{gathered}$ CI on Difference | Coverage (\%) | $\underset{(p \text {-value })}{\text { Tstat }}$ | $\underset{(p-\text { value })}{S-W^{1} \text { Stat }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Bootstrap ${ }^{2}$ | Mean | Lower | $\begin{aligned} & -5.731 \mathrm{E}-03 \\ & (1.337 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-3.207 \mathrm{E}-02, \\ 2.061 \mathrm{E}-02) \end{gathered}$ | 72.62 | $\begin{gathered} -7.370 \mathrm{E}-01 \\ (0.548) \end{gathered}$ | $\begin{gathered} 0.1465 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{aligned} & \hline-1.284 \mathrm{E}-03 \\ & (9.898 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-2.078 \mathrm{E}-02, \\ 1.821 \mathrm{E}-02) \end{gathered}$ | 78.14 | $\begin{gathered} -6.444 \mathrm{E}-02 \\ (0.437) \end{gathered}$ | $\begin{gathered} 0.1512 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.273 \mathrm{E}-12 \\ & (2.882 \mathrm{E}-14) \end{aligned}$ | $\begin{gathered} (-9.337 \mathrm{E}-12, \\ -9.195 \mathrm{E}-12) \end{gathered}$ | >99.99 | $\begin{gathered} -3.130 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.189 \mathrm{E}-12 \\ (2.096 \mathrm{E}-14) \end{gathered}$ | $\begin{aligned} & (9.150 \mathrm{E}-12, \\ & 9.239 \mathrm{E}-12) \end{aligned}$ | >99.99 | $\begin{gathered} 4.444 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
| Jackknife ${ }^{3}$ | Mean | Lower | $\begin{aligned} & -5.733 \mathrm{E}-03 \\ & (5.366 \mathrm{E}-04) \end{aligned}$ | $\begin{aligned} & (-6.790 \mathrm{E}-03, \\ & -4.677 \mathrm{E}-03) \end{aligned}$ | >99.99 | $\begin{gathered} -1.069 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0793 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{aligned} & \hline-1.274 \mathrm{E}-03 \\ & (5.080 \mathrm{E}-04) \end{aligned}$ | $\begin{aligned} & (-2.274 \mathrm{E}-03, \\ & -2.733 \mathrm{E}-04) \end{aligned}$ | 99.33 | $\begin{gathered} \hline-2.508 \mathrm{E}+00 \\ (0.013) \end{gathered}$ | $\begin{gathered} 0.0915 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.273 \mathrm{E}-12 \\ & (5.741 \mathrm{E}-03) \end{aligned}$ | $\begin{gathered} (-1.231 \mathrm{E}-02, \\ 1.231 \mathrm{E}-02) \end{gathered}$ | 50.00 | $\begin{gathered} -1.619 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.189 \mathrm{E}-12 \\ (1.292 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-2.771 \mathrm{E}-03, \\ 2.771 \mathrm{E}-03) \end{gathered}$ | 50.00 | $\begin{gathered} 7.024 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ | N/A |
| Partial Batch ${ }^{4}$ | Mean | Lower | $\begin{aligned} & -5.733 \mathrm{E}-03 \\ & (1.536 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-3.598 \mathrm{E}-02, \\ 2.451 \mathrm{E}-02) \end{gathered}$ | 64.54 | $\begin{gathered} -3.734 \mathrm{E}-01 \\ (0.709) \end{gathered}$ | $\begin{gathered} 0.0793 \\ (<0.0001) \end{gathered}$ |
|  |  | Upper | $\begin{aligned} & -1.274 \mathrm{E}-03 \\ & (1.300 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-2.688 \mathrm{E}-02, \\ 2.433 \mathrm{E}-02) \end{gathered}$ | 53.90 | $\begin{gathered} -9.801 \mathrm{E}-02 \\ (0.922) \end{gathered}$ | $\begin{gathered} 0.0916 \\ (<0.0001) \end{gathered}$ |
|  | Median | Lower | $\begin{aligned} & -9.273 \mathrm{E}-12 \\ & (3.144 \mathrm{E}-14) \end{aligned}$ | $\begin{gathered} (-9.342 \mathrm{E}-12, \\ -9.195 \mathrm{E}-12) \end{gathered}$ | >99.99 | $\begin{gathered} -2.889 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |
|  |  | Upper | $\begin{gathered} 9.189 \mathrm{E}-12 \\ (1.847 \mathrm{E}-14) \end{gathered}$ | $\begin{gathered} (9.150 \mathrm{E}-12, \\ 9.229 \mathrm{E}-12) \end{gathered}$ | >99.99 | $\begin{gathered} 4.974 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ | N/A |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas; ${ }^{4}$ Program 32.sas

The above analyses were based on the rates of AE occurrence. The following analyses were intended to represent subgroup analyses on these AEs. Specifically, AE locations were compared on the characteristics severity, relation to treatment, and seriousness. Data patterns 1-4 assumed various levels of these characteristics. Analysis was performed with the 2-PL EX model based on Gibbs sampling iterations 11,500 to 50,000 with the lags $\boldsymbol{\theta}=695, \boldsymbol{A}=365$, and $\boldsymbol{B}=430$ for paired treatment groups with $n=$

250 patients. Results were generated with the nonparametric bootstrap estimation approach for the median of the spline approximation to the exact AUC.

## Analyses on AE Characteristics

The previous analyses were based on fixed historical rate data. The following sections repeated this analysis, but for a different type of data. Instead of being concerned with the occurrence of AEs as a totality, this section focuses on AE subgroup analyses. Specifically, these analyses were constructed to determine if IRT is able to correctly characterize the relations between treatment groups on characteristics of the AEs. With this characterization, a more thorough understanding of safety can be developed by physicians.

Four data patterns of AE characteristics were investigated, and these characteristics were arbitrarily classified as locations. As the characteristics increased in the magnitude of medical consequences, the location of the AE increased as well. These locations were provided in Table 4 (p. 84). Data Pattern 1 represented non-serious AEs that were not related to treatment and were non-severe. For both treatment groups, the AE was assigned a location of 1 . Data Pattern 2 represented non-serious AEs that were possibly related to treatment and were severe for Treatment A and non-serious AEs that were doubtfully related to treatment and were non-severe for Treatment B. The location was 3 and 6, respectively, for treatments A and B. Data Pattern 3 represented serious AEs that were definitely related to treatment and were severe for Treatment A and non-serious AEs that were not related to treatment and were non-severe for Treatment B. The location
was 1 and 20, respectively, for treatments A and B. Data Pattern 4 consisted of randomly allocated locations for both treatment groups.

Data Pattern 1 (Same Characteristics). For treatments A and B, AE Location 1 was simulated to approximately occur $20 \%$ and $10 \%$, respectively, of the time. This simulation was designed so that Treatment A was visually different from and not statistically equivalent $(\Delta=10 \%)$ to Treatment B in AE Predisposition $(\boldsymbol{\theta})$. Figure 55 presents the IRF for treatments A and B. For Treatment A, the spline approximation (bias) was $5.127(-0.093)$ for the exact area under the IRF (5.220). For Treatment B, the spline approximation (bias) was $3.854(-0.037)$ for the exact area under the IRF (3.891).


## Figure 55. PPI Plot for Data Pattern 1 AE Location 1

Table 35 (p. 175) presents the results for evaluating superiority and equivalence for AE Location 1. Results demonstrated that AE Predisposition ( $\boldsymbol{\theta}$ ) was statistically
different between treatments $(p$-value $=0.018)$. For the equivalence objective, the results on the lower effect $(p$-value $=0.109)$ and upper effect $(p$-value $=0.005)$ did not result in statistically equivalent treatment groups for a margin of $10 \%$ with $95 \%$ confidence (twosided).

Table 35
Superiority and Equivalence Analysis: Data Pattern 1 AE Type 1

| Type of <br> Analysis | Effect | Difference <br> $(\boldsymbol{S E})$ | 95\% (2-sided) <br> CI on Difference | Coverage <br> $(\%)$ | Tstat <br> $(\boldsymbol{p}$-value) $)$ | S-W <br> $(\boldsymbol{p}$-value $)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Superiority | N/A | $4.107 \mathrm{E}-04$ | $(8.300 \mathrm{E}-05$, | 99.11 | $2.685 \mathrm{E}+00$ | 0.1741 <br> $(0.018)$ |
|  |  | $(1.418 \mathrm{E}-04)$ | $7.120 \mathrm{E}-04)$ |  | $(0.0001)$ |  |
| Equivalence | Lower | $3.327 \mathrm{E}-04$ | $(-7.017 \mathrm{E}-05$, | 94.56 | $1.706 \mathrm{E}+00$ <br> $(0.109)$ | 0.1768 <br> $(<0.0001)$ |
|  |  | $(1.539 \mathrm{E}-04)$ | $6.505 \mathrm{E}-04)$ |  | $3.312 \mathrm{E}+00$ <br>  <br>  | Upper |
|  | $4.482 \mathrm{E}-04$ |  |  |  |  |  |
| $(1.373 \mathrm{E}-04)$ | $(1.633 \mathrm{E}-04$, | 99.74 | 0.1640 <br> $(<0.0001)$ |  |  |  |

${ }^{1}$ Shapiro-Wilk test for normality of the effect.

Data Pattern 2 (Comparable Characteristics). For Treatment B, AE Location 3 was simulated with approximate occurrence $20 \%$. For Treatment A, AE Location 6 was simulated with approximate occurrence $10 \%$. The occurrence of all other AE locations was $0 \%$. This simulation was designed so that Treatment A was visually different from and not statistically equivalent $(\Delta=10 \%)$ to Treatment B in AE Predisposition $(\boldsymbol{\theta})$ for AE Locations 3 and 6.

Figures 56-57 (p. 177) present the IRF for AE Locations 3 and 6. For Treatment A and AE Location 3, the spline approximation (bias) was $5.128(-0.095)$ for the exact area under the IRF (5.22337). For Treatment B, the spline approximation (bias) was 5.095 $(-0.089)$ for the exact area under the IRF (5.184). For Treatment A and AE Location 6,
the spline approximation (bias) was $4.014(-0.039)$ for the exact area under the IRF (4.053). For Treatment B, the spline approximation (bias) was $5.080(-0.091)$ for the exact area under the IRF (5.170).

Tables 36-37 (p. 178) present the results for evaluating superiority and equivalence for AE Locations 3 and 6. For AE Location 3, the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter was not statistically different between treatments $(p$-value $=0.491)$, nor statistically equivalent for $\Delta=10 \%$ ( $p$-value for lower effect $=0.141 ; p$-value for upper effect $=0.497$ ). For AE Location 6, AE Predisposition $(\boldsymbol{\theta})$ was not statistically different between treatments $(p$-value $=0.184)$, nor statistically equivalent at $\Delta=10 \%$ ( $p$-value for lower effect $=0.062 ; p$-value for upper effect $=0.426$ ). The results for AE Location 3 and 6 were unexpected. Exploratory analysis revealed that lack of statistical power resulted in the insignificance for these superiority analyses. This means that a sample size of $n=250$ was too small to demonstrate a statistically significant difference between treatment groups for these data.

Data Pattern 3 (Extreme Characteristics). For Treatment B, AE Location 1 was simulated with approximate occurrence $5 \%$. For Treatment A, AE Location 20 was simulated with approximate occurrence $20 \%$. The occurrence of all other AE locations was $0 \%$. The simulation was designed so that Treatment A was visually different from and not statistically equivalent $(\Delta=10 \%)$ to Treatment B in AE Predisposition $(\boldsymbol{\theta})$ for AE Locations 1 and 20.


Figure 56. PPI Plot for Data Pattern 2 AE Location 3


Figure 57. PPI Plot for Data Pattern 2 AE Location 6

Table 36
Superiority and Equivalence Analysis: Data Pattern 2 AE Location 3

| Type of <br> Analysis $^{2}$ | Effect | Difference <br> $(\boldsymbol{S E} \boldsymbol{)}$ | 95\% (2-sided) <br> CI on Difference | Coverage <br> $(\%)$ | Tstat <br> $(\boldsymbol{p}$-value) $)$ | S-W <br> $(\boldsymbol{p}$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Superiority | N/A | $-1.601 \mathrm{E}-05$ | $(-2.043 \mathrm{E}-04$, | 75.46 | $-1.843 \mathrm{E}-01$ | 0.2616 |
|  |  | $(9.134 \mathrm{E}-05)$ | $1.660 \mathrm{E}-04)$ |  | $(0.491)$ | $(0.0001)$ |
| Equivalence | Lower | $-1.338 \mathrm{E}-04$ | $(-3.291 \mathrm{E}-04$, | 92.93 | $-1.556 \mathrm{E}+00$ | 0.2537 |
|  |  | $(9.430 \mathrm{E}-05)$ | $5.449 \mathrm{E}-05)$ |  | $(0.141)$ | $(0.0001)$ |
|  | Upper | $5.683 \mathrm{E}-05$ | $(-1.262 \mathrm{E}-04$, | 75.16 | $6.350 \mathrm{E}-01$ | 0.2377 |
|  |  | $(9.088 \mathrm{E}-05)$ | $2.438 \mathrm{E}-04)$ |  | $(0.497)$ | $(0.0001)$ |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas

Table 37
Superiority and Equivalence Analysis: Data Pattern 2 AE Location 6

| Type of Analysis ${ }^{2}$ | Effect | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { CI on Difference } \end{gathered}$ | Coverage (\%) | Tstat ( $p$-value) | S-W ${ }^{1}$ Stat ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Superiority | N/A | $\begin{aligned} & -1.653 \mathrm{E}-04 \\ & (1.052 \mathrm{E}-04) \end{aligned}$ | $\begin{gathered} (-3.605 \mathrm{E}-04, \\ 8.751 \mathrm{E}-05) \end{gathered}$ | 90.81 | $\begin{gathered} -1.394 \mathrm{E}+00 \\ (0.184) \end{gathered}$ | $\begin{gathered} 0.1798 \\ (0.0001) \end{gathered}$ |
| Equivalence | Lower | $\begin{aligned} & -2.487 \mathrm{E}-04 \\ & (1.085 \mathrm{E}-04) \end{aligned}$ | $\begin{gathered} (-4.512 \mathrm{E}-04 \\ 1.407 \mathrm{E}-05) \end{gathered}$ | 96.91 | $\begin{gathered} -2.031 \mathrm{E}+00 \\ (0.062) \end{gathered}$ | $\begin{gathered} 0.1736 \\ (0.0001) \end{gathered}$ |
|  | Upper | $\begin{aligned} & -6.916 \mathrm{E}-06 \\ & (1.000 \mathrm{E}-04) \end{aligned}$ | $\begin{gathered} (-2.380 \mathrm{E}-04 \\ 1.874 \mathrm{E}-04) \end{gathered}$ | 78.72 | $\begin{gathered} -6.628 \mathrm{E}-02 \\ (0.426) \end{gathered}$ | $\begin{gathered} 0.1890 \\ (0.0001) \end{gathered}$ |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas

Figures 58-59 (p. 179) present the IRF for AE Locations 1 and 20. For Treatment A and AE Location 1, the spline approximation (bias) was $5.125(-0.086)$ for the exact area under the IRF (5.212). For Treatment B, the spline approximation (bias) was 3.884 $(-0.044)$ for the exact area under the IRF (3.928). For Treatment A and AE Location 20, the spline approximation (bias) was $5.139(-0.092)$ for the exact area under the IRF


Figure 58. PPI Plot for Data Pattern 3 AE Location 1


Figure 59. PPI Plot for Data Pattern 3 AE Location 20
(5.231). For Treatment B, the spline approximation (bias) was 5.164 ( -0.091 ) for the exact area under the IRF (5.255).

Table 38 (below) and Table 39 (p. 181) present the results for evaluating superiority and equivalence for AE Location 1 and 20, respectively. For AE Location 1, AE Predisposition $(\boldsymbol{\theta})$ was not statistically different between treatment groups ( $p$-value $=$ 0.161 ), nor statistically equivalent for $\Delta=10 \%$ (lower effect $p$-value $=0.402$; upper effect $p$-value $=0.098$ ). For AE Location 20, AE Predisposition $(\boldsymbol{\theta})$ was not statistically different between treatment groups ( $p$-value $=0.535$ ), nor statistically equivalent for $\Delta=10 \%$ (lower effect $p$-value $=0.269$; upper effect $p$-value $=0.135$ ). For AE Location 1, it was evident from Figure 58 that the treatment groups were different, and lack of significance was a function of statistical power. The results for AE Location 20 were unexpected. Exploratory analysis did not reveal any anomalies in the performed analysis, and further investigation may be warranted.

## Table 38

Superiority and Equivalence Analysis: Data Pattern 3 AE Location 1

| Type of Analysis ${ }^{2}$ | Effect | Difference (SE) | $\begin{gathered} \text { 95\% (2-sided) } \\ \text { CI on Difference } \end{gathered}$ | Coverage (\%) | Tstat (p-value) | S-W ${ }^{1}$ Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Superiority | N/A | $\begin{gathered} 2.194 \mathrm{E}-04 \\ (1.502 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-1.052 \mathrm{E}-04 \\ 5.249 \mathrm{E}-04) \end{gathered}$ | 91.95 | $\begin{gathered} 1.457 \mathrm{E}+00 \\ (0.161) \end{gathered}$ | $\begin{gathered} 0.1748 \\ (0.0001) \end{gathered}$ |
| Equivalence | Lower | $\begin{gathered} 8.633 \mathrm{E}-05 \\ (1.454 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-2.077 \mathrm{E}-04 \\ 3.903 \mathrm{E}-04) \end{gathered}$ | 79.91 | $\begin{gathered} 5.779 \mathrm{E}-01 \\ (0.402) \end{gathered}$ | $\begin{gathered} 0.1815 \\ (0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 2.698 \mathrm{E}-04 \\ (1.574 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-6.337 \mathrm{E}-05, \\ 6.513 \mathrm{E}-04) \end{gathered}$ | 95.10 | $\begin{gathered} 1.772 \mathrm{E}+00 \\ (0.098) \end{gathered}$ | $\begin{gathered} 0.1669 \\ (0.0001) \end{gathered}$ |

[^2]Table 39
Superiority and Equivalence Analysis: Data Pattern 3 AE Location 20

| Type of Analysis ${ }^{2}$ | Effect | Difference (SE) | 95\% (2-sided) CI on Difference | Coverage (\%) | Tstat ( $p$-value) | S- $\mathbf{W}^{1}$ Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Superiority | N/A | $\begin{gathered} 1.763 \mathrm{E}-07 \\ (8.981 \mathrm{E}-05) \end{gathered}$ | $\begin{gathered} (-1.990 \mathrm{E}-04 \\ 1.812 \mathrm{E}-04) \end{gathered}$ | 73.27 | $\begin{gathered} 1.675 \mathrm{E}-03 \\ (0.535) \end{gathered}$ | $\begin{gathered} 0.2911 \\ (0.0001) \end{gathered}$ |
| Equivalence | Lower | $\begin{aligned} & -1.516 \mathrm{E}-04 \\ & (1.336 \mathrm{E}-04) \end{aligned}$ | $\begin{gathered} (-4.780 \mathrm{E}-04, \\ 1.358 \mathrm{E}-04) \end{gathered}$ | 86.56 | $\begin{aligned} & -1.112 \\ & (0.269) \end{aligned}$ | $\begin{gathered} 0.2940 \\ (0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 1.180 \mathrm{E}-04 \\ (7.143 \mathrm{E}-05) \end{gathered}$ | $\begin{gathered} (-3.930 \mathrm{E}-05, \\ 2.587 \mathrm{E}-04) \end{gathered}$ | 93.24 | $\begin{gathered} 1.585 \mathrm{E}+00 \\ (0.135) \end{gathered}$ | $\begin{gathered} 0.2744 \\ (0.0001) \end{gathered}$ |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30 .sas

Data Pattern 4 (Random Characteristics). Simulations resulted in AE Locations 1-20 having occurrence $15.6 \%$ to $22.0 \%$ for Treatment A and $14.0 \%$ to $22.0 \%$ for Treatment B. The simulations were constructed to mimic an unsafe medical product that would unlikely be approved for market use by a regulatory agency such as the Food and Drug Administration.

Figure 60 (p. 182) presents the IRF across AE Locations 1-20. For Treatment A, the spline approximation (bias) was $3.413(-0.025)$ for the exact area under the overall IRF (3.438). For Treatment B, the spline approximation (bias) was 3.412 ( -0.0158 ) for the exact area under the overall IRF (3.427).

Table 40 (p. 182) presents the results for evaluating superiority and equivalence across all AE locations. Results demonstrated that AE Predisposition ( $\boldsymbol{\theta})$ was not statistically different between treatment groups ( $p$-value $=0.343$ ), nor statistically equivalent for $\Delta=10 \%$ (lower effect $p$-value $=0.286$; upper effect $p$-value $=0.190$ ).

Figure 60 reveals that an equivalence assumption between treatment groups may be more
reasonable than superiority. Regardless, the equivalence analysis did not have sufficient power for $n=250$ at $95 \%$ confidence (two-sided).


Figure 60. PPI Plot for Data Pattern 4 Across All AE Locations

Table 40
Superiority and Equivalence Analysis: Data Pattern 4 All AE Locations

| Type of Analysis ${ }^{2}$ | Effect | Difference (SE) | $\begin{aligned} & \text { 95\% (2-sided) } \\ & \text { CI on Difference } \end{aligned}$ | Coverage (\%) | Tstat ( $p$-value) | S-W ${ }^{1}$ Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Superiority | N/A | $\begin{gathered} 1.193 \mathrm{E}-04 \\ (1.592 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-2.327 \mathrm{E}-04 \\ 3.961 \mathrm{E}-04) \end{gathered}$ | 82.83 | $\begin{gathered} 7.200 \mathrm{E}-01 \\ (0.343) \end{gathered}$ | $\begin{gathered} 0.4278 \\ (0.0001) \end{gathered}$ |
| Equivalence | Lower | $\begin{aligned} & -2.246 \mathrm{E}-04 \\ & (1.872 \mathrm{E}-04) \end{aligned}$ | $\begin{gathered} (-5.778 \mathrm{E}-04, \\ 2.521 \mathrm{E}-04) \end{gathered}$ | 85.71 | $\begin{gathered} -9.884 \mathrm{E}-01 \\ (0.286) \end{gathered}$ | $\begin{gathered} 0.5300 \\ (0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 3.064 \mathrm{E}-04 \\ (1.949 \mathrm{E}-04) \end{gathered}$ | $\begin{gathered} (-1.553 \mathrm{E}-04 \\ 6.980 \mathrm{E}-04) \end{gathered}$ | 90.50 | $\begin{gathered} 1.372 \mathrm{E}+00 \\ (0.190) \end{gathered}$ | $\begin{gathered} 0.3374 \\ (0.0001) \end{gathered}$ |

${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30 .sas

In summary, exploratory analyses did not statistically demonstrate superiority for Data Patterns 2-3 and equivalence for Data Pattern 4. The sample size of $n=250$ patients did not have sufficient power for substantiating study objectives. In order to avoid such an issue in future studies, the obtained results could be used as historical data for properly powering such a study. This exercise would result in the minimum number of patients required to demonstrate either superiority or equivalence between the treatment groups. Furthermore, the results from investigating the 7 equivalence margin definitions (Table 30, p. 166) should be considered for equivalence studies. This study arbitrarily used a margin of $10 \%$. Other studies may require a different equivalence margin.

## Research Question 3 Results

This section compares superiority and equivalence sample size results between the available existing methods for paired binomial endpoints and the power function derived by the author on IRT results. Sample size results were based on the simulated AE data assuming historical rates.

## Existing Methodology Superiority Power Analysis

Each Monte Carlo simulation resulted in a $2 \times 2$ contingency table for $N=500,000$ patients. The average cell count (proportion) of AEs that occurred for both treatment groups was 642.14 ( 0.001284287 ). The average cell count (proportion) of AEs that occurred for Treatment A and not Treatment B was 10.7134 (0.00002143). The average cell count (proportion) of AEs that occurred for Treatment B and not Treatment A was 11.1368 ( 0.00002227 ). The average cell count (proportion) of AEs that did not occur for
both treatment groups was 499336.01 ( 0.99867 ). Table 41 presents the superiority sample sizes for the existing methods that are based on the difference or ratio of discordant probabilities. These sample size requirements ranged from 174,451 to $522,548,365$.

Table 41
Superiority Sample Size Requirements for Existing Methods

| Method $^{1}$ | Computed Power | Observed Type I <br> Error | Required Sample Size |
| :---: | :---: | :---: | :---: |
| Existing Method 1 | 0.80000 | 0.025001 | $478,108,748$ |
| Existing Method 2 | 0.80007 | 0.025000 | 174,451 |
| Existing Method 3 | 0.80000 | 0.025001 | $478,108,748$ |
| Existing Method 4 | 0.80000 | 0.025001 | $522,548,365$ |
| Existing Method 5 | 0.80000 | 0.025001 | $478,054,794$ |
| Existing Method 6 | 0.80000 | 0.025001 | $478,149,214$ |
| Existing Method 7 | 0.80000 | 0.025001 | $478,108,748$ |
| Existing Method 8 | 0.80000 | 0.025001 | $478,108,750$ |

${ }^{1}$ Program 34.sas

IRT Superiority Power Analysis

According to Table 27 (p. 161), the nonparametric bootstrap approach based on the spline approximation to the exact AUC resulted in a median $(S E)$ of $1.088 \times 10^{-13}$ $\left(2.163 \times 10^{-13}\right)$. For the superiority IRT power function derived by the author in Chapter III, the non-central parameter and critical value were computed as 7.87049 and 3.85063 , respectively (Program 35.sas). With these values and the nonparametric bootstrap results, a minimum of 933 patients were required to demonstrate that the treatment groups were statistically different for $97.5 \%$ confidence (one-sided) and at least $80 \%$ power. The analyses for this study were based on $n=30$ and 250 .

## Existing Methodology Equivalence Power Analysis

For the existing approaches for computing a sample size for an equivalence study objective using paired treatment groups, the effect was based on the difference in discordant probabilities $\left(\mathrm{p}_{10}-\mathrm{p}_{01}\right)$. For the 1,000 simulated AE datasets, the average of $\mathrm{p}_{10}$ and $\mathrm{p}_{01}$, respectively, was 0.000021427 and 0.000022274 . Using these results, the first existing method (Wald approach) required 95 patients for demonstrating equivalence for a $10 \%$ equivalence margin between treatment groups with $95 \%$ confidence (two-sided) and at least $80 \%$ power. The second existing method (RMLE approach) required 73,300 patients for demonstrating equivalence for a $10 \%$ equivalence margin between treatment groups with $95 \%$ confidence (two-sided) and at least $80 \%$ power (Program 34.sas).

## IRT Equivalence Power Analysis

According to Table 32 (p. 169), the nonparametric bootstrap approach based on the spline approximation to the exact AUC resulted in a median $(S E)$ of $-9.162 \times 10^{-12}$ $\left(3.654 \times 10^{-12}\right)$ for the lower effect and $9.294 \times 10^{-12}\left(4.597 \times 10^{-12}\right)$ for the upper effect. For the equivalence IRT power function derived by the author in Chapter III, the noncentral parameter and critical value were computed as 12.3644 and 4.00687 , respectively, for the lower effect. These values were 8.1750 and 4.00398 , respectively, for the upper effect (Program 35.sas). With these values and the nonparametric bootstrap results, 59 patients (lower effect) and 60 patients (upper effect) were required to demonstrate that the treatment groups were statistically equivalent for $95 \%$ confidence (two-sided) and at least $80 \%$ power.

## Summary of Results

This study investigated three research questions surrounding the use of Item Response Theory (IRT) for modeling rare binary events. The first research question pertained to whether or not IRT could be used to sufficiently model this type of data. Bayesian IRT models were assessed for multiple chains, convergence of parameter means and standard deviations, autocorrelation, and goodness-of-fit (GoF). The second research question pertained to determining the best statistical inference for evaluating superiority and equivalence study objectives. For each approximation to the exact area under paired Item Response Functions (IRFs), the standard error (SE), bias, and coverage were investigated. The last research question pertained to investigating sample size requirements for rare binary events. The Bayesian IRT approach derived by the author was compared to existing methods.

## Research Question 1

The first section of this study investigated the adequacy of three IRT approaches for modeling rare binary event data. These data were illustrated in terms of transfusionrelated adverse events (AEs). These models assumed different probability density functions (PDFs) of the discrimination parameter. Monte Carlo simulations $(1,000)$ were used to investigate nine types of AEs. Each analysis was performed with simple random samples (SRS) of sizes $n=30$ and 250 patients.

Multiple Chains of Starting Values. Three different chains of starting values were investigated to determine if the study results were an artifact of the data. If these starting
values resulted in substantial differences between these IRT parameters, IRT would not be a sufficient approach for modeling rare binary events. Study results demonstrated that the mean and $S D$ of these parameters were comparable between the three IRT models, and this consistency was not substantially affected by sample size. In conclusion, the 2-PL, 2-PL EX, and 2-PL MEX IRT models were capable of modeling rare binary events.

Stationary State. Assessment of the stationary state and rate of estimator convergence was the next step for evaluating the IRT models under investigation. For the 2-PL IRT model, across the three chains of starting values, the mean of the AE Predisposition $(\boldsymbol{\theta})$, discrimination, and difficulty parameters converged from 25,000 to 30,000 Gibbs sampler iterations (GSI) for $n=30$ and from 10,000 to 25,000 GSI for $n=$ 250. The $S D$ of the IRT parameters converged from 10,000 to 25,000 GSI for $n=30$ and from 30,000 to 35,000 GSI for $n=250$. For the 2-PL EX IRT model, across the three chains of starting values, the mean of the IRT parameters converged from 15,000 to 30,000 GSI for $n=30$ and from 10,000 to 25,000 GSI for $n=250$. The $S D$ of the IRT parameters converged from 10,000 to 20,000 GSI for $n=30$ and from 15,000 to 20,000 GSI for $n=250$. For the 2-PL MEX IRT model, across the three chains of starting values, the mean of the discrimination and difficulty parameters converged for 7,000 GSI for $n=30$ and 6,000 GSI for $n=250$. The AE Predisposition $(\boldsymbol{\theta})$ parameter did not converge for $n=30$ and 250 . The $S D$ of the discrimination and difficulty parameters converged from 6,000 to 7,000 GSI for $n=30$ and for 7,000 for $n=250$. The mean and $S D$ of the AE Predisposition ( $\boldsymbol{\theta}$ ) parameter did not converge for $n=30$ and 250 .

Summary. This section demonstrated that both the 2-PL and 2-PL EX IRT models achieved stationary states. The latter IRT model had a faster rate of convergence than the
former model. For the 2-PL MEX IRT model, some of the estimators failed to converge. This lack of convergence may have occurred due to a combination of computational limitations and the sampling algorithm efficiency of the Gibbs sampler. As a result, this model was removed as a viable option for modeling rare binary event data. Furthermore, because all three chains converged to the same location, although at a different rate, the remaining results in this section were limited to the chain of starting values $\boldsymbol{A}^{(i)}=\mathbf{- 1}$ and $\boldsymbol{\theta}^{(i)}=\mathbf{- 1}$.

Autocorrelation. The next step in evaluating the 2-PL and 2-PL EX IRT models involved identifying the lag function that removed serial autocorrelation data from the stationary state. The lag function was computed for the AE Predisposition ( $\boldsymbol{\theta})$, discrimination, and difficulty parameters for sample sizes of $n=30$ and 250. For the 2-PL IRT model, the maximum lag of the discrimination parameter was 185 for $n=30$ and 350 for $n=250$. The maximum lag of the difficulty parameter was 90 for $n=30$ and 350 for $n=250$. The lag function of the AE Predisposition $(\boldsymbol{\theta})$ parameter was 330 for $n=30$ and 565 for $n=250$. For the 2-PL EX IRT model, the maximum lag of the discrimination parameter was 95 for $n=30$ and 365 for $n=250$. The maximum lag of the difficulty parameter was 210 for $n=30$ and 430 for $n=250$. The lag function of the AE Predisposition $(\boldsymbol{\theta})$ parameter was 180 for $n=30$ and 695 for $n=250$.

Goodness-of-Fit. GoF was the last Bayesian IRT model assumption evaluated. This study evaluated GoF with posterior probability interval (PPI) plots, residual plots, and the asymptotic chi-square Bock's Index (BI). PPI plots did not exhibit and exhibited patient discriminatory power for the 2-PL and 2-PL EX IRT models, respectively. The discrimination or slope parameter was consistently close to zero for 2-PL and above 0.75
for 2-PL EX. The residual plots demonstrated relatively small errors for both of these models. Next, all IRFs were monotonic for 2-PL EX. Last, as would be predicted by Toribio (2006), the BI results were inconsistent with the PPI plots. As the discriminatory power of AE Predisposition $(\boldsymbol{\theta})$ increased, the BI also increased. This statistic would have proven to be useful if it demonstrated a negative correlation with patient discriminatory power. As a result, the BI has limited value for IRT GoF of rare binary event data. In summary, the 2-PL IRT model did not demonstrate sufficient discriminatory power (i.e., discrimination $<0.75$ for all AE types). For the 2-PL EX IRT model, discrimination was greater than 0.75 for all AE types. As a result, this model was recommended for statistical inference and power analysis application.

## Research Question 2

The second section of this study investigated statistical inference and its components for the three IRT models. Analysis was performed on simulated data for historical AE rates and four different patterns of AE characteristics. Discussion of results in this section was limited to the former type of data modeled with 2-PL EX.

Linear trapezoid and spline approximations of the exact area under paired IRFs were computed for each parametric and nonparametric estimator. The bootstrap, jackknife, and partial batch approaches were used to compute the statistical inference of these effects for superiority and equivalence study objectives. These components were $S E$, bias, test statistic, $p$-value, confidence interval, and coverage.

Area Under IRFs. IRFs for treatments A and B were simulated to be comparable for historical AE rates. The median linear trapezoid and spline approximations to the
exact area under IRFs were $4.688(-0.221)$ and $4.841(-0.069)$, respectively, across treatment groups and sample sizes. These results demonstrated that the spline approach more closely approximated the exact area than the linear trapezoid approximation. As a result, the remainder of this section was limited to a discussion of statistical inference based on the spline approximation.

The bootstrap, jackknife, and partial batch estimation approaches were used to perform statistical inference on the parametric and nonparametric estimators. These estimators were compared for superiority and equivalence study objectives.

Superiority Analyses. Shapiro-Wilk results consistently demonstrated that the distribution of the paired effects was not statistically normal. As a result, statistical inference was limited to nonparametric analysis in this section. Next, adequate coverage was achieved for the bootstrap and partial batch estimation approaches for $n=30$ and 250. A claim of superiority was not statistically substantiated for both the bootstrap and partial-batch estimation approaches. This result was expected because the IRFs were simulated to be comparable. The CIs on the median of the paired difference contained zero, but the sample size $n=250$ resulted in a significant $p$-value . This abnormal result was verified to be a function of sample size in conjunction with the test statistic.

Equivalence Analyses. Shapiro-Wilk results consistently demonstrated that the distribution of the lower and upper equivalence paired effects was not statistically normal. As a result, statistical inference was limited to nonparametric analysis in this section. Next, adequate coverage was achieved for the bootstrap estimation approach. For the sample size $n=30$, this approach did not result in the statistical substantiation of equivalence for $\Delta=10 \%$ at $95 \%$ confidence (two-sided). Exploratory analysis found that
this result was due to insufficient statistical power. For $n=250$, this approach resulted in the statistical substantiation of equivalence for the same margin and confidence level. This result was expected because the IRFs were simulated to be comparable.

## Research Question 3

The last section of this study was concerned with computing sample size requirements on simulated paired binomial endpoints for superiority and equivalence study objectives. IRT sample size approaches derived by the author were compared to the available existing methods. Sample size estimation was performed for the nonparametric bootstrap estimation approach.

Superiority. The bootstrap approach for the median required a minimum of 933 patients for two paired treatment groups with $95 \%$ confidence (two-sided) and at least $80 \%$ power. The discordance rate of 0.000044 resulted in a minimum sample size requirement of 174,451 for two paired treatment groups with $95 \%$ confidence (two-sided) and at least $80 \%$ power. All of these approaches achieved the target Type I error. These large sample sizes were expected because the IRFs were simulated to be comparable. Designing a study that is powered to statistically differentiate treatment groups that are comparable is rarely financially feasible. Equivalence was demonstrated to be a viable solution to this problem.

Equivalence. The nonparametric bootstrap approach for the median required a minimum of 60 patients for paired treatment groups for a $10 \%$ equivalence margin with $95 \%$ confidence (two-sided) and at least $80 \%$ power. The discordance rate of 0.000044 resulted in a minimum sample size requirement of 95 for two paired treatment groups
with the same margin, confidence level, and power. All of these approaches achieved the target Type I error.

Equivalence Margin. As previously discussed, the $10 \%$ equivalence margin was arbitrarily chosen. If real historical data were available prior to the start of this study, a more exact margin would have been computed and utilized. Given that such data were unavailable, a section was dedicated to equivalence margin computation for rare and more frequent AEs. Researchers should consult this section for designing equivalence studies. The scientific validity of the computed equivalence margin should then be assessed.

## CHAPTER V

## DISCUSSION AND CONCLUSIONS

## Discussion

This study investigated various data scenarios common to the transfusion medicine industry. Specifically, regulatory authorities across the world are very conscientious and conservative about the transfusion of blood products such as platelets, red blood cells, and plasma. Various companies are tasked with developing medical devices and technology that work as intended and pose minimal safety risk to the patient. The responsibility of regulatory authorities is protection of the public by making sure that such risk is truly minimal.

The difficulty for both of these parties pertains to their ability to accurately quantify the precision of such risk. The first step in this decision-making process is quantifying acceptable patient risk. With this definition, trials can then be properly designed and powered so that clinically relevant hypotheses about patient risk may be statistically substantiated. This process may improve the rigor and generalizability of safety conclusions to patient populations and decisions from medical stakeholders.

One of the key issues that complicates these inferential safety conclusions pertains to data type. The transfusion of blood products infrequently results in adverse events (AEs). That is, the occurrence of AEs is very low in blood transfusion settings. Although these events are rare, the transfusion-related AEs that do occur typically have serious
medical consequences for the patient. As a result, being able to make reliable conclusions and decisions regarding the transfusion of blood products may be benefited with the rigor of inferential treatment comparisons in AE occurrence.

Existing methods are available for designing studies for these comparisons, but these methods have one common limitation. They require sample sizes that are not financially feasible in a rare event setting. Furthermore, reliable and monitored clinical trial data are expensive, and may cost upwards of $\$ 10,000-\$ 12,000$ per patient. As a result, the concept of statistically substantiated safety of rare events is desired but not currently practical.

The purpose of this study was to determine if a different type of statistical approach presented a viable solution to this financial dilemma. This study appeared to be the first study to investigate AE hypotheses with the psychometric approach Item Response Theory (IRT) in an industry clinical trial (ICT) setting. This study developed the statistical inference and power analysis capabilities of this approach for the paired or $k$-sample matched study design. Hypothesis testing was based on the comparing the area under IRFs between paired treatment groups for superiority and equivalence study objectives.

Rare binary event data were also used in terms of a larger scope of application. These data were used as a worst-case scenario to stress the limits of IRT capability. Results demonstrated that when a gamma probability density function was used to model the discrimination parameter, the applicable IRT model (2-PL EX) reliably modeled these data. As a result, statistical inference and power functions based on the IRT model results
were trustworthy. This means that this study developed a new approach for designing studies on planned treatment comparisons of most rare binary event variables.

## DIF versus EIF

In the ICT setting, these objectives are known as superiority and equivalence. Hypotheses for these objectives are designed to demonstrate minimum and maximum effects, respectively. In other words, superiority is used to demonstrate that groups are statistically different in a response variable when such a difference is expected. When it is desired to demonstrate that groups behave comparably, equivalence is recommended.

Superiority has been commonly employed in IRT application, but it typically has been referred to as Differential Item Functioning (DIF). DIF has been used to determine if groups (e.g., males and females) are statistically different on latent trait parameters. The findings from these analyses may then be used to appropriately modify or improve a test instrument. In an ICT, DIF or superiority may be used to determine if a new treatment is statistically better than another group in an efficacy parameter (e.g., plasma hemoglobin).

Clinical trial statisticians may argue that DIF or superiority is only one of two key types of objectives needed for comparing treatment groups on response variables. In some study settings, a placebo may not be available, legal, or moral. In these cases, a new treatment is typically compared to an existing treatment with an objective other than superiority. For this comparison, the objective is to demonstrate that a new treatment is comparable to an existing treatment on a response variable. If this comparability is statistically substantiated, the treatment groups are said to be equivalent.

Unfortunately, there is some confusion about these two objectives in the literature, and is evident in the testing of hypotheses for evaluating data distributions. This confusion stems from interpreting a nonsignificant $p$-value from a superiority analysis as to indicate that the groups are equivalent. In terms of data distributions, nonsignificant $p$-values represent the desired data distribution.

This logic is incorrect for two key reasons. The effects for these two objectives are computed differently. This means that one objective cannot be the inverse of the other. Second, the structure of the hypothesis sets is different. Superiority can be based on a single one-sided hypothesis or a single two-sided hypothesis. Equivalence is always based on two one-sided hypotheses.

Given this new information, equivalence appears to be a new tool in IRT application, and the author defined it as the Equivalence Item Function (EIF). DIF is recommended for use when groups are expected to differ on latent trait parameters. EIF is recommended for use when groups are expected to be comparable on these parameters. As a result, access to both DIF and EIF may provide IRT researchers and statisticians a more complete set of tools for comparing groups on scientifically relevant parameters. This, in turn, may lead to more reliable study decisions. Clinical trials that compare a new treatment to both an existing treatment and a placebo is an example of the need for both DIF and EIF in a single study.

## Advantages of Study

In addition to the development of EIF, this study was valuable in several respects. The methodological tools developed in this study have wide application outside of the
study parameters. The statistical inference and power functions derived by the author are not limited to the paired study design. Although this study performed analyses on paired treatment groups, the statistical inference and power functions are readily adaptable to independent treatment groups. The researcher would simply need to appropriately modify the covariances and degrees of freedom for the confidence intervals and test statistics. Next, the IRT statistical inference and power functions developed in this study are universal to all types of data concerned with a response function. This is the case because the statistic used for analysis was the area under Item Response Functions. These functions are commonly used in IRT application. Third, the latent trait was assumed to be unidimensional. The developed statistical inference and power functions were invariant to this data structure. In multidimensional data structures, the same analysis could be performed within each trait, or performed as an aggregation of traits. Last, this study focused on the two-parametric logistic IRT model because of its ability to differentiate patients on AE Predisposition ( $\boldsymbol{\theta})$. In that most if not all IRT methodology follows a comparable path, the developed statistical inference and power functions are not limited to any specific type of IRT model that is concerned with a response function.

## New IRT Knowledge Base

This study investigated rare binary event data with three Bayesian forms of the two-parameter logistic IRT model. These data were characterized in terms of transfusionrelated AEs. All of these models assumed a normal distribution for the latent trait $\boldsymbol{Z}, \mathrm{AE}$ Predisposition $(\boldsymbol{\theta})$, and difficulty $(\boldsymbol{B})$ parameters. The 2-PL model (common approach) assumed a standard normal density for the discrimination parameter ( $\boldsymbol{A}$ ). The 2-PL EX
model (exchangeable approach) assumed a gamma density for $\boldsymbol{A}$. The 2-PL MEX model (mixed-exchangeable approach) assumed a mixture of two normal densities for $\boldsymbol{A}$, and used the beta PDF for removing outliers from analysis.

The empirical distribution of the incidence of AEs readily followed gamma and beta densities. 2-PL failed because of its normality requirement for $\boldsymbol{A}$. Parameter convergence was not consistent for 2-PL MEX, and this lack of convergence may have resulted from computational limitations due to the inefficiency of the Gibbs sampling algorithm. As anticipated and supported, 2-PL EX IRT adequately modeled rare binary event data.

After a determination was made as to which IRT model, if any, was adequate for modeling rare binary event data, statistical inference and power analysis was investigated for 2-PL EX. The area-under-the-curve (AUC) statistic was used for these analyses, because of its nonparametric property. AUC can be equally computed for known and unknown data distributions. In IRT application, many inferential decisions are based on the IRF, which may or may not have known form. As a result, the AUC statistic was assumed to be sufficient for statistical inference and power analysis on the area under IRFs.

Two AUC approximations were compared to the exact area under paired IRFs. Spline approximations were found to be slightly closer to the exact areas than the linear trapezoid approximation. This finding was expected because the spline is more robust for nonlinear functions and optimal for cubic polynomials. This result means that statistical inference based on the spline approximation will be more reliable, and sample size requirements will be less. The investigation of the spline was considered especially
important in regards to IRT application. In order for test instruments to be of value, they need to be able to differentiate respondents on latent trait parameters. The greater the patient discriminatory power of the IRFs, the more nonlinear the IRF. For the logistic IRT model, the IRF is theoretically cubic.

Last, statistical analysis was performed using parametric and nonparametric estimators using the bootstrap, jackknife, and partial batch approaches. The nonparametric estimator (i.e., median) for the bootstrap approach resulted in consistently better properties than the mean. The former finding was expected in that the distribution of the paired effects was non-normal for all analyses. This conclusion was based on the standard errors being consistently smaller and the coverage being consistently higher for the median.

The author did not have any preconceived expectations about these three approaches. The study found that the bootstrap approach resulted in better statistical inference properties than the jackknife and partial batch approaches. As a result, the nonparametric bootstrap approach was recommended for statistical inference. As a result, decisions regarding AUC comparisons based on findings from this combination of approaches were considered reliable because the inference correctly characterized superiority and equivalence data scenarios.

This methodology was then incorporated into the derived IRT power functions. Two important factors were investigated to understand the sensitivity of these functions. Simulations were performed to intentionally mimic superiority and equivalence data scenarios. Sample size requirements for superiority consistently increased as the two IRFs approximated each other. The same trend occurred for equivalence when the IRFs were
simulated to be different. These results demonstrated that the derived IRT power functions were sound.

The IRT sample size requirements were then compared to existing methods for superiority and equivalence study objectives. The IRT sample sizes for both of these objectives were consistently smaller than the existing methods. More importantly, the IRT approach for equivalence resulted in a sample size that may be financially feasible in an ICT. This sample size requirement may be further reduced by taking advantage of findings from the research topics recommended in the next section.

Directions for Future Research

## Data Pattern Anomalies

This study investigated 2-PL EX IRT model results for four patterns of AE characteristics. Two of the simulations were designed so that IRFs would be noticeably different between treatment groups. Treatment A possessed a higher rate of AE occurrence, and vice versa for Treatment B. An unexplained anomaly occurred for two of these analyses. When simulating a superiority data scenario where the occurrence of Treatment B (Control group) was larger than Treatment A (Test group), the IRFs showed notable separation. But when this rate of occurrence was reversed, the IRFs did not separate. They unexpectedly represented an equivalence data scenario. Exploratory analysis was performed, and no sources of this anomaly could be identified. Future research pertaining to the sensitivity of the 2-PL EX IRT model is warranted.

## Goodness-of-Fit

Inferential methods that correctly characterize goodness-of-fit (GoF) is a weakness of Bayesian IRT methodology. Future methods that are a function of the predicted and observed values associated with patient discriminatory power may prove to be valuable. As this power increases while residuals remain relatively small, the test statistic for the GoF method should theoretically decrease. The development of such a statistic would enable more accurate conclusions regarding posterior probability and residual plots.

## Bayesian Sampling Algorithms

This study used the inverse cumulative distribution function (CDF) transformation method for the Bayesian sampling algorithm. This algorithm was used to generate sequences of Gibbs samplers. The AE Predisposition ( $\boldsymbol{\theta})$, discrimination, and difficulty parameters converged slowly for the 2-PL and 2-PL EX IRT models. This convergence did not consistently occur for 2-PL MEX.

The 2-PL MEX findings may have been a result of computational limitations from the sampling algorithm. Future research may demonstrate that other sampling algorithms such as Metropolis-Hastings (Johnson \& Albert, 1999), Adaptive Rejection Sampling (Gilks \& Wild, 1992), and Slice Sampling (Neal, 2003) converge faster for 2-PL and 2-PL EX, and converge sufficiently for 2-PL MEX.

## Nonlinear Measurement Error

The ability to correctly diagnose an AE could play a role in statistical inference findings and sample size requirements. Several studies have demonstrated that the reliability of these diagnoses may not always be high (Cowell, Dawid, Hutchinson, \& Spiegelhalter, 1991; Hutchison, 1986, 1991; Hutchison, Dawid, Spiegelhalter, Cowell, \& Roden, 1991; Kane-Gill, Kirisci, \& Pathak, 2005; Lanctôt, 1989; Lanctôt, Kwok, \& Naranjo, 1995; Macdeo, Marques, Ribeiro, \& Teixeira, 2005; Naranjo et al., 1981; Pere, Begaud, Haramburu, \& Albin, 1986). Ignoring this form of measurement error could potentially confound statistical inference findings and sample size requirements. If this measurement error could be accounted for in the IRT model, then this contamination could be minimized or even prevented. ICTs have a solution for addressing this concern with reliable AE diagnosis.

In ICTs, study physicians record the occurrence of any adverse events experienced by the patient on a Case Report Form. These physicians may either be blinded to or aware of the treatment they administered to their patient. A Data Safety Monitoring Board (DSMB) may then be used to evaluate the accuracy of the diagnoses from the study physician. The DSMB commonly consists of a panel of physicians who are knowledgeable about the study indication, independent of the study, and blinded to patient treatment (Dixon, 2008). These parameters are in place to minimize concerns regarding decision bias. Over the course of an ICT, the DSMB adjudicates each reported AE. The study database may then contain the AE diagnosed by the study physician as well as the DSMB. With both of these data sources available, the reliability of AE
diagnosis can be computed and incorporated into estimation. For the logistic IRT model investigated in this study, reliability would be characterized as nonlinear measurement error. Newly developed methods are available for integrating this type of error into Bayesian nonlinear mixed regression models (Mallick, 2011; Sinha, Mallick, Kipnis, \& Carroll, 2010).

## Concluding Remarks

This study began with a discussion of the requirements regulatory authorities such as the FDA use to approve medical products for market use. Product efficacy must be statistically substantiated, but this requirement is infrequent for AEs. This study attempted to present a viable solution to this problem. If this study was able to develop such a solution, the pharmaceutical industry would have the ability to appropriately power their ICTs on both efficacy and AE hypotheses. Also, regulatory authorities, such as the FDA, would have access to a new way of thinking about investigating product safety in terms of AEs and other safety parameters. Being able to change an industry standard may prove to be very valuable to the community.

This ability may allow the pharmaceutical industry to generalize clinical safety findings on a small sample of patients to target populations. This may mean that a medical product's safety profile would be more thoroughly understood before its release to the market, not thereafter. Furthermore, a tremendous advantage of this study was that the derived group-level statistical inference and power analysis was readily applicable to the patient in terms of the Patient Response Function. With the advent of Risk Prediction Algorithms, this methodology may be a viable option for differentiating and identifying
patients who are most predisposed ("at-risk") to experiencing a particular type of AE.
Known and reliable probabilities that correctly characterize clinically relevant predisposition may help physicians customize or modify treatment regimens so that the AE effects are minimized or negated all together.

## REFERENCES

Agresti, A., \& Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in $2 \times 2$ contingency tables. Biometrics, 61, 515-523.

Alexandrowicz, R. (2002). Zur bestimmung der effektgrößen bein Likelihood-Ratio-Test nach Anderson [Power analysis for Anderson's likelihood ratio test]. Unpublished doctoral dissertation, University of Vienna, Austria.

Altham, P. M. E. (1971). The analysis of matched proportions. Biometrika, 58, 561-576.
Amarnani, R. (2009). Two theories, one theta: A gentle introduction to item response theory as an alternative to classical test theory. The International Journal of Educational and Psychological Assessment, 3, 104-109.

American Red Cross. Retrieved December 12, 2010, from http://www.redcrossblood.org/learn-about-blood/blood-components

Bailer, A. J. (1988). Testing for the equality of area under the curves when using destructive measurement techniques. Journal of Pharmacokinetics and Biopharmaceutics, 16, 303-309.

Bandos, A. I., Rockette, H. E., \& Gur, D. (2007). Exact bootstrap variances of the area under ROC curve. Communications in Statistics - Theory and Methods, 36, 24432461.

Bandyopadhyay, P. S., \& Forster, M. R. (2010). Philosophy of statistics: An introduction. In P. S. Bandyopadhyay \& M. R. Forster (Eds.), Handbook of the philosophy of statistics. North Holland: Elsevier.

Bartolucci, F., \& Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. Psychometrika, 70, 31-43.

Bartoszynski, R., \& Niewiadomska-Bugaj, M. (1996). Probability and statistical inference. New York: Wiley.

Battles, J. B., Kaplan, H. S., Van der Schaaf, T. W., \& Shea, C. E. (1998). The attributes of medical event-reporting systems: Experience with a prototype medical eventreporting system for transfusion medicine. Archives of Pathology \& Laboratory Medicine, 122, 231-238.

Bauer, P. (2005). Therapeutic equivalence - Some comments. Biometrical Journal, 47, 99-103.

Bernardo, J. M. (2003). Bayesian statistics. In R. Viertl (Ed.) Probability and statistics of the Encyclopedia of Life Support Systems (EOLSS). Oxford, UK: UNESCO.

Bishop, Y. M. M., Feinberg, S. E., \& Holland, P. W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Bjorner, J. B., Kosinski, M., \& Ware, J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test ( $\left.\mathrm{HIT}^{\mathrm{TM}}\right)$. Quality of Life Research, 12, 913-933.

Black, S. A., Markides, K. S., \& Ray, L. A. (2003). Depression predicts increased incidence of adverse health outcomes in older Mexican Americans with type 2 diabetes. Diabetes Care, 26, 2822-2828.

Blackwelder, W. C. (2002). Showing a treatment is good because it is not bad: When does "noninferiority" imply effectiveness? Controlled Clinical Trials, 23, 52-54.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock, R. D. (1997). A brief history of item response theory. Educational Measurement: Issues and Practice, 16, 21-33.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. Journal of Educational Measurement, 37, 307-327.

Bonate, P. L. (1998). Coverage and precision of confidence intervals for area under the curve using parametric and non-parametric methods in a toxicokinetics experimental design. Pharmaceutical Research, 15, 405-410.

Bremer, R., Perez, J., Smith, P., \& Westfall, P. (2004, November). Grid computing at Texas Tech University using SAS. Paper presented at the Proceedings of the 14th Annual South-Central SAS User's Group Regional Conference, Austin, TX.

Bridgewater, B., Neve, H., Moat, N., Hooper, T., \& Jones, M. (1998). Predicting operative risk for coronary artery surgery in the United Kingdom: A comparison of various risk prediction algorithms. Heart, 79, 350-355.

Castelloe, J. M. (2000). Sample size computations and power analysis with the $S A S^{\circledR}$ system. Paper 265-25. Conference SUGI 25. April 9-12, 2000, Indiana Convention Center, Indianapolis, Indiana.

Caufmann, E., \& MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument: Identifying race and gender differential item functioning among juvenile offenders. Educational and Psychological Measurement, 66, 502-521.

Centers for Disease Control and Prevention. (2010). The National Healthcare Safety Network (NHSN) manual. Retrieved from http://www.cdc.gov/nhsn/PDFs/ hemovigModuleProtocol_current.pdf

Cepicka, L. (2003). The Rasch model in the motor ability testing. Journal of Human Kinetics, 10, 99-106.

Chhibber, V., \& King, K. E. (2010). Management of the therapeutic apheresis patient. In B. C. McLeod, Z. M. Szczepiorkowski, R. Weinstein, \& J. L. Winters, J. L. (Eds.), Apheresis: Principles and practice (3rd ed.). Bethesda, MD: AABB Press.

Chiou, W. L. (1978). Critical evaluation of the potential error in pharmacokinetic studies of using the linear trapezoidal rule method for the calculation of the area under the plasma level-time curve. Journal of Pharmacokinetics and Biopharmaceutics, 6, 539-546.

Chow, S.-C., \& Shao, J. (2006). On non-inferiority margin and statistical tests in active control trials. Statistics in Medicine, 25, 1101-1113.

Chow, S.-C., Shao, J., \& Wang, H. (2008). Sample size calculations in clinical research ( $2^{\text {nd }}$ ed.). Boca Raton, FL: Chapman \& Hall/CRC.

Cole, J. C., Lin, P., \& Rupnow, M. F. T. (2009). Minimal important differences in the Migraine-Specific Quality of Life Questionnaire (MSQ) version 2.1. Cephalalgia, 29, 1180-1187.

Connett, J. E., Smith, J. A., \& McHugh, R. B. (1987). Sample size and power for pairmatched case-control studies. Statistics in Medicine, 6, 53-59.

Connor, R. J. (1987). Sample size for testing differences in proportions for the pairedsample design. Biometrics, 43, 207-211.

Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C.-H., Crane, P. K., Gibbons, L. E., Hays, R. D., McHorney, C. A., Ocepek-Welikson, K., Raczek, A. E., Teresi, J. A., \& Reeve, B. B. (2007). IRT health outcomes data analysis project: An overview and summary. Quality of Life Research, 16, 121-132.

Cooper, J. A., Miller, G. J., \& Humphries, S. E. (2005). A comparison of the PROCAM and Framington point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. Atherosclerosis, 181, 93100.

Cowell, R. G., Dawid, A. P., Hutchinson, T. A., \& Spiegelhalter, D. J. (1991). A Bayesian expert system for the analysis of adverse drug reactions. Artificial Intelligence in Medicine, 3, 257-270.

Cowen, P. S., \& Moorhead, S. (2006). Current issues in nursing (7 ${ }^{\text {th }}$ ed.). St. Louis, MO: Mosby.

Crane, P. K., Gibbons, L. E., Narasimhalu, K., Lai, J.-S., \& Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. Quality of Life Research, 16, 101-114.

D’Agostino, R. B., Grundy, S., Sullivan, L. M., \& Wilson, P. (2001). Validation of the Framington coronary heart disease prediction scores: Results of a multiple ethnic groups investigation. The Journal of the American Medical Association, 286, 180197.

Daly, L. (1992). Simple SAS macros for the calculation of exact binomial and Poisson confidence limits. Computers in Biology and Medicine, 22, 351-361.
de Ayala, R. J. (2009). The theory and practice of item response theory. New York: The Guilford Press.
de Gruijter, D. N. M. (2004). Comparison of the nonparametric Mokken model and parametric IRT models using latent class analysis. Applied Psychological Measurement, 18, 27-34.
de Leeuw, E. D., \& Hox, J. J. (1998). Detecting aberrant response patterns in multi-item scales: Some nonparametric indices and a computer program. Bulletin de Méthodologie Sociologique, 58, 68-78.

Department of Veterans Affairs. (2008). Disclosure of adverse events to patients (VHA Directive 2008-002). Washington, DC: Veterans Health Administration.

Dixon, D. O. (2008). Data and Safety Monitoring Boards: Introduction to the principles and practice of clinical research biostatistics research branch: NIAID. January 14, 2008.

Douglas, J. A. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. Journal of Educational Measurement, 33, 465-484.

Drachler, L., Marshall, T., \& de Carvalho, L. J. C. (2007). A continuous-scale measure of child development for population-based epidemiological surveys: A preliminary study using item response theory for the Denver Test. Paediatric and Perinatal Epidemiology, 21, 138-153.

Drugs. (2011). Retrieved March 28, 2011, from http://www.drugs.com/aspirin.html

Duffy, S. W. (1984). Asymptotic and exact power for the McNemar Test and its analogue with R controls per case. Biometrics, 40, 1005-1015.

DuMouchel, W. (2010). Multivariate Bayesian logistic regression for clinical safety data. $4^{\text {th }}$ Seattle Symposium in Biostatistics: Clinical Trials, November 23, 2010.

Dunning, A. M. (2007). Comparing bootstrap and jackknife variance estimation methods for area under the ROC curve using one-stage cluster survey data. Unpublished master's thesis, Virginia Commonwealth University.

Dupont, W. D. (1988). Power calculations for matched case-control studies. Biometrics, 44, 1157-1168.

Eaves, L., Erkanli, A., Silberg, J., Angold, A., Maes, H. H., \& Foley, D. (2005). Application of Bayesian inference using Gibbs sampling to item-response theory modeling of multi-symptom genetic data. Behavior Genetics, 35, 765-780.

Eisen, S. V., Wilcox, M., Leff, H. S., Schaefer, E., \& Culhane, M. A. (1999). Assessing behavioral health outcomes in outpatient programs: Reliability and validity of the BASIS-32. The Journal of Behavioral Health Services \& Research, 26, 5-17.

Embretson, S. E., \& Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Empey, P. E. (2010). Genetic predisposition to adverse drug reactions in the intensive care unit. Critical Care Medicine, 38, S106-S116.

Fergusson, D., Hebert, P., \& Shapiro S. (2002). The before/after study design in transfusion medicine: Methodologic considerations. Transfusion Medicine Review, 16, 296-303.

Finch, H., \& Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. Applied Psychological Measurement, 31, 292-307.

Food and Drug Administration (FDA). (1999). Guidance for industry: Q6B specifications: Test procedures and acceptance criteria for biotechnological/biological products. Retrieved from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryIn formation/Guidances/UCM073488.pdf

Fox, J.-P., \& Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika, 66, 271-288.

Froelich, A. G., \& Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. Applied Psychological Measurement, 32, 138-155.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, \& D. J. Spiegelhalter (Eds.), Markov chain Monte Carlo in practice (pp. 131-144). London: Chapman and Hall.

Geraci, J. M., Rosen, A. K., Ash, A. S., McNiff, K. J., \& Moskowitz, M. A. (1993). Predicting the occurrence of adverse events after coronary artery bypass surgery. Annals of Internal Medicine, 118, 18-24.

Gilks, W. R., \& Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41, 337-348.

Glas, C. A. W., \& Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. Applied Psychological Measurement, 27, 217-233.

Glickman, M. E., Seal, P., \& Eisen, S. V. (2009). A non-parametric Bayesian diagnostic for detecting differential item functioning in IRT models. Health Services \& Outcomes Research Methodology, 9, 145-161.

Goethals, R. (1994). Die prakitsche erprobung von alternativen zur multiple-choicevorgabe bei computertests [Experiences with alternatives for traditional multiplechoice response formats in computerized testing]. Unpublished doctoral dissertation, University of Vienna, Austria.

Grattagliano, I., Portincasa, P., Mastronardi, M., Palmieri, V. O., \& Palasciano, G. (2005). Esomeprazole-induced central fever with severe myalgia. The Annals of Pharmacotherapy, 39, 757-760.

Graves, S. (1978). On the Neyman-Pearson theory of testing. The British Journal for the Philosophy of Science, 29, 1-23.

Hambleton, R. K., \& Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14(2), 75-96

Hambleton, R. K. \& Cook, L. L. (1997). Latent Trait Models and Their Use in the Analysis of Educational Test Data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K., \& van der Linden, W. K. (1982). Advances in item response theory and applications: An introduction. Applied Psychological Measurement, 6, 373378.

Hanley, J. A., \& Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. Academic Radiology, 4, 49-58.

Hanley, J. A., \& McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29-36.

Hanson, S. A., \& France, C. R. (2004). Predonation water ingestion attenuates negative reactions to blood donation. Transfusion, 44, 924-928.

Haq, I. U., Jackson, P. R., Yeo, W. W., \& Ramsay, L. E. (1995). Sheffield risk and treatment table for cholesterol lowering for primary prevention of coronary heart disease. Lancet, 346, 1467-1471.

Hardouin, J.-B. (2007). Rasch analysis: Estimation and tests with Raschtest. The Stata Journal, 7, 1-23.

Hauschke, D. (2001). Choice of delta: A special case. Drug Information Journal, 35, 875-879.

Haynie, W. J. (2007). Effects of test taking on retention learning in technology education: A meta-analysis. Journal of Technology Education, 18, 24-36.

Hays, R. D., Morales, L. S., \& Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Medical Care, 38, II28-II42.

Heddle, N. M., Klama, L. N., Griffith, L. R., Shukla, G., \& Kelton, J. G. (1993). A prospective study to identify the risk factors associated with acute reactions to platelet and red cell transfusions. Transfusion, 33, 794-797.

Hedeker, D., Mermelstein, R. J., \& Flay, B. R. (2006). Application of IRT models for intensive longitudinal data. In T. A. Walls, \& J. L. Schafer (Eds.), Models for intensive longitudinal data (pp. 84-108). New York: Oxford University Press.

Heyse, J. F., \& Stine, R. (2000, June). Use of the overlapping coefficient for measuring the similarity of treatments. Paper presented at the American Statistical Association 2000 Proceedings, Park City, UT.

Higgins, T. L., Estafanous, F. G., Loop, F. D., Beck, G. J., Blum, J. M., \& Paranandi, L. (1992). Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients: A clinical severity score. JAMA, 267, 2344-2348.

Hoehler, F. K. (1995). Exact power calculations when the dependent variable is a single proportion and the number of events is small. Computers in Biology and Medicine, 25(5), 471-479.

Hogg, R. V., \& Craig, A. T. (1995). Introduction to mathematical statistics (5th ed.). New York: Macmillan.

Hogg, R. V., McKean, J. W., \& Craig, A. T. (2005). Introduction to mathematical statistics (6th ed.). New Jersey: Prentice Hall.

Holland, P. W., \& Thayer, D. T. (1986, April). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the $67^{\text {th }}$ annual meeting of the American Educational Research Association, San Francisco.

Holman, R., Math, M., Glas, C. A. W., \& de Haan, R. J. (2003). Power analysis in randomized clinical trials based on item response theory. Controlled Clinical Trials, 24, 390-410.

Hshieh, P., \& Ng, T.-H. (2007, July). Noninferiority testing with a given percentage of the control as the noninferiority margin. Paper presented at the Joint Statistical Meetings, Salt Lake City, UT.

Huh, Y. O., \& Lichtiger, B. (1995, January-April). Nonhemolytic transfusion reaction to platelet concentrates: Induction by cytokines released during storage. Current Issues in Transfusion Medicine, 4.

Hung, H. M. J., Wang, S. J., Tsong Y., Lawrence, J., \& O’Neil, R. T. (2003). Some fundamental issues with non-inferiority testing in active controlled trials. Statistics in Medicine, 22, 213-225.

Hutchison, T. A. (1986). A Bayesian approach to assessment of adverse drug reactions: Evaluation of a care of acute renal failure. Drug Information Journal, 20, 475-482.

Hutchison, T. A. (1991). Bayesian ADE assessment. Drug Information Journal, 25, 235241.

Hutchison, T. A., Dawid, A. P., Spiegelhalter, D. J., Cowell, R. G., \& Roden, S. (1991). Computer aids for probabilistic assessment of drug safety. I: A spreadsheet program. Drug Information Journal, 25, 29-39.

Iannelli, V. (2010). Flu shot reactions and flu shot side effects. Retrieved April 10, 2011, from http://pediatrics.about.com/od/weeklyquestion/a/05_flu_shot_rxn.htm

International Conference on Harmonization. (1994). The extent of population exposure to assess clinical safety for drugs intended for long-term treatment of non-lifethreatening conditions (ICH E1). Geneva, Switzerland: Author.

International Conference on Harmonization. (1995). Clinical safety data management: Definitions and standard for expedited reporting (ICH E2a). Geneva, Switzerland: Author.

International Conference on Harmonization. (1995). Structure and content of clinical study reports (ICH E3). Geneva, Switzerland: Author.

International Conference on Harmonization. (1996). Guidance for industry E6 good clinical practice: Consolidated guidance (ICH E6). Geneva, Switzerland: Author.

International Conference on Harmonization. (1997). General considerations for clinical trials (ICH E8). Geneva, Switzerland: Author.

International Conference on Harmonization. (1998). Statistical principles for clinical trials (ICH E9). Geneva, Switzerland: Author.

International Conference on Harmonization. (2000). Clinical investigation of medicinal products in the pediatric population (ICH E11). Geneva, Switzerland: Author.

International Conference on Harmonization. (2000). Choice of control group in clinical trials (ICH E10). Geneva, Switzerland: Author.

Jagsi, R., Kitch, B. T., Weinstein, D. F., Campbell, E. G., Hutter, M., \& Weissman, J. S. (2005). Residents report on adverse events and their causes. Archives of Internal Medicine, 165, 2607-2613.

Jannarone, R. J., Yu, K. F., \& Laughlin, J. E. (1990). Easy Bayes estimation for Raschtype models. Psychometrika, 55, 449-460.

Johnson, M. S., \& Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. Applied Psychological Measurement, 29, 369-400.

Johnson, R. E., \& Liu, H. (2000, April). Pseudo-random numbers: Out of uniform. Paper presented at SUGI 25, Indianapolis, IN.

Johnson, V. E., \& Albert, J. H. (1999). Ordinal data modeling. New York: Springer.
Jones, L. V. (1952). Tests of hypotheses: One-sided vs. two-sided alternatives. Psychological Bulletin, 49, 43-46.

Junker, B. W. (1997, August). From psychometrics through item response theory to complex statistical analyses. Paper presented at Invited Paper Meeting \#69: An Overview of Contemporary Psychometrics, 51st Session of the International Statistical Institute, Istanbul, Turkey.

Kane-Gill, S. L., Kirisci, L., \& Pathak, D. S. (2005). Are the Narajno criteria reliable and valid for determination of adverse drug reactions in the intensive care unit? The Annals of Pharmacotherapy, 39, 1823-1827.

Khalid, S., Usman, M., \& Khurshid, M. (2010). Acute transfusion reactions encountered in patients at a tertiary care center. Journal of the Pakistan Medical Association, 60, 832-836.

Kleinman, S., Chan, P., \& Robillard, P. (2003). Risks associated with transfusion of cellular blood components in Canada. Transfusion Medicine Reviews, 17, 120-162.

Kok, F. G., Mellenbergh, G. J., \& van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22, 295-303.

Koroto, T. (2009). An approximate likelihood ratio method for testing equality of two dependent proportions. Asian Journal of Mathematics and Statistics, 2, 15-19.

Krediet, C. T. P., van Dijk, N., Linzer, M., van Lieshout, J. J., \& Weiling, W. (2002). Management of vasovagal syncope: Controlling or aborting faints by leg crossing and muscle tensing. Circulation, 106, 1684-1689.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model - Some critical suggestions on traditional approaches. Journal of International Testing, 5, 377-394.

Laake, J. H., \& Røttingen, J. A. (2001). Rocuronium and anaphylaxis - A statistical challenge. Scandinavian Society of Anaesthesiology and Intensive Care Medicine, 45, 1196-1203.

Lachenbruch, P. A. (1992). On the sample size for studies based upon McNemar's Test. Statistics in Medicine, 11, 1521-1525.

Lachenbruch, P. A., \& Lynch, C. J. (1998). Assessing screening tests: Extensions of McNemar's Test. Statistics in Medicine, 17, 2207-2217.

Lachin, J. M. (1992). Power and sample size evaluation for the McNemar Test with application to matched case-control studies. Statistics in Medicine, 11, 1239-1251.

Lanctôt, K. (1989). Using microcomputers to simplify the Bayesian causality assessment of adverse drug reactions. Pharmaceutical Medicine, 4, 185-195.

Lanctôt, K., Kwok, M. C. O., \& Naranjo, C. A. (1995). Computerized Bayesian evaluation of adverse events. Drug Information Journal, 29, 319-325.

Lange, S., \& Freitag, G. (2005). Choice of delta: Requirements and reality - Results of a systematic review. Biometrical Journal, 47, 12-27.

Lee, M. T., \& Lusher, J. M. (1991). The problem of therapeutic equivalence with paired qualitative data: An example from a clinical trial using haemophiliacs with an inhibitor to factor VIII. Statistics in Medicine, 10, 443-441.

Lee, P. M. (1997). Bayesian statistics: An introduction. New York: John Wiley \& Sons.
Lewis, C. (2006). Note on unconditional and conditional hypothesis testing: A discussion of an issue raised by van der Linden and Sotaridona. Journal of Educational and Behavioral Statistics, 31, 305-309.

Lin, C. T., Hosmane, B. S., Olson, P. J., \& Padley, R. J. (2001). Analysis of adverse events in titration studies. The Journal of Statistical Planning and Inference, 96, 129-142.

Linacre, J. M. (1994). Sample size and item calibration stability. Rasch Measurement Transactions, 7, 328.

Liu, J. P., Hsueh, H.-M., Hsieh, E., \& Chen, J. J. (2002). Tests for equivalence or noninferiority for paired binary data. Statistics in Medicine, 21, 231-245.

Lloyd, C. J. (1990). Confidence intervals from the difference between two correlated proportions. Journal of the American Statistical Association, 85, 1154-1158.

Lohr, S. (1999). Sampling: Design and analysis. Pacific Grove, CA: Duxbury Press.
Lord, F. M. (1953). An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. Psychometrika, 18, 57-75.

Lord, F. M., \& Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lu, Y., \& Bean, J. A. (1995). On the sample size for one-sided equivalence of sensitivities based upon McNemar's Test. Statistics in Medicine, 14, 1831-1839.

Lui, K.-J. (2005). Sample size calculation for testing non-inferiority and equivalence under Poisson distribution. Statistical Methodology, 2, 37-48.

Lynch, S. M. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. New York: Springer.

Macdeo, A. F., Marques, F. B., Ribeiro, C. F., \& Teixeira, F. (2005). Causality assessment of adverse drug reactions: Comparison of the results obtained from published decisional algorithms and from the evaluations of an expert panel. Pharmacoepidemiology and Drug Safety, 14, 885-890.

Mallick, B. (2011, July). Semiparametric Bayesian approach for measurement error models. Paper presented at the Joint Statistical Meetings, Miami, FL.

May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. Journal of Educational and Behavioral Statistics, 31, 63-79.

May, H., Cole, R., Haimson, J. \& Perez-Johnson, I. (2011). Assessing the conditional reliability of state assessments. Society for Research on Educational Effectiveness Conference. (ERIC Document Reproduction Service No. ED512708)

May, W. L., \& Johnson, W. D. (1997). Confidence intervals for differences in correlated binary proportions. Statistics in Medicine, 16, 2127-2136.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus nonparametric IRT modeling of Likert-type personality data. Multivariate Behavioral Research, 40, 261-279.

The Mayo Clinic. (2011). Retrieved May 15, 2011, from http://www.mayoclinic.com/ health/strep-throat/DS00260/DSECTION=symptoms

McKinley, R. L., \& Way, W. D. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models. TOEFL Technical Report TR-5 RR-92-16. (ERIC Document Reproduction Service No. ED385583)

Meijer, R. R., \& Sijtsma, K. (2001). Methodology review: Evaluating person fit. Applied Psychological Measurement, 25, 107-135.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Miettinen, O. S. (1968). The matched pairs design in the case of all-or-none responses. Biometrics, 24, 339-352.

Mitra, S. K. (1958). On the limiting power function of the frequency chi-square test. Annals of Mathematical Statistics, 29, 1221-1233.

Nam, J.-M. (1997). Establishing equivalence of two treatments and sample size requirements in matched-pairs design. Biometrics, 53, 1422-1430.

Nandakumar, R., \& Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics, 18, 41-68.

Naranjo, C. A., Busto, U., Sellers, E. M., Sandor, P., Ruiz, I., Roberts, E. A., Janecek, E., Domecq, C., \& Greenblatt, D. J. (1981). Naranjo: A method for estimating the probability of adverse drug reactions. Clinical Pharmacology and Therapeutics, 30, 239-245.

Navarro-Fontestad, M. C., González-Álvarez, I., Fernández-Terul, C., Bermejo, M., \& Casabó, V. G. (2005). Methods for estimating the AUC and its variability in sparse sampling designs. Paper presented at the Sociedad Espaňola De Farmacia Industrial Y Galénica.

Neal, R. M. (2003). Slice sampling. Annals of Statistics, 31, 705-767.

Nedelman, J. R., Gibiansky, E., \& Lau, D. T. W. (1995). Applying Bailer's method for AUC confidence intervals to sparse sampling. Pharmaceutical Research, 12, 124128.

Ness, P. M., Shirey, R. S., Thoman, S. K., \& Buck, S.A. (1990). The differentiation of delayed serologic and delayed hemolytic transfusion reactions: Incidence, long-term serologic findings, and clinical significance. Transfusion, 30, 688-693.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. Statistics in Medicine, 17, 2635-2650.

Newman, B. H., \& Roth, A. J. (2005). Estimating the probability of a blood donation adverse event based on 1000 interviewed whole-blood donors. Transfusion, 45, 1715-1721.

Neyman, J. (1977). Frequentist probability and frequentist statistics. Synthese, 36, 97131.

Ng, T.-H. (2001). Choice of delta in equivalence testing. Drug Information Journal, 35, 1517-1527.

Nishtala, P. S., Fois, R. A., McLachlan, A. J., Bell, J. S., Kelly, P. J., \& Chen, T. F. (2009). Anticholinergic activity of commonly prescribed medications and neuropsychiatric adverse events in older people. Journal of Clinical Pharmacology, 49, 1176-1184.

Noel, Y., \& Dauvier, B. (2007). A beta item response model for continuous bounded responses. Applied Psychological Measurement, 31, 147-173.

Nova Scotia Department of Health. (2005). Identification and management of a transfusion reaction. Nova Scotia, Canada: Author.

Oberg, K. C. (1999). Adverse drug reactions. American Journal of Pharmaceutical Education, 63.

Olive, D. J. (2005). A simple confidence interval for the median. Retrieved July 1, 2010, from http://www.math.siu.edu/olive/ppmedci.pdf

Patz, R. J., \& Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics, 24, 342-366.

Patz, R. J., \& Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics, 24, 146-178.

Pearson, E. S., \& Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the non-central f-distribution. Biometrika, 38, 112-130.

Pere, J. C., Begaud, B., Haramburu, F., \& Albin, H. (1986). Computerized comparison of six adverse drug reaction assessment procedures. Clinical Pharmacology and Therapeutics, 40, 451-461.

Perkins, A. J., Stump, T. E., Monahan, P. O., \& McHorney, C. A. (2006). Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. Quality of Life Research, 15, 331-348.

Pham-Gia, T., \& Turkkan, N. (2008) Bayesian analysis of a $2 \times 2$ contingency table with dependent proportions and exact sample size. Statistics, 42, 127-147.

Reeve, B. B., \& Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers \& R. Hays (Eds.), Assessing quality of life in clinical trials: Methods and practice (2 ${ }^{\text {nd }}$ ed.) (pp. 5573). Oxford, UK: Oxford University Press.

Rochester, C. G. (2003, September). Regulatory statistical perspectives on safety issues in drug development. Paper presented at the FDA/Industry Statistics Workshop, Rockville, MD.

Rochester, C. G. (2009, March). The quantitative safety analysis plan. Paper presented at the Drug Information Association Annual Euromeeting, Berlin, Germany.

Ross, S. M. (1997). Introduction to probability models. New York: Academic Press.
Rudner, L. M. (1998). An on-line, interactive, computer adaptive testing mini-tutorial. Retrieved from http://edresearch.org/scripts/cat

Rühl, H., Bein, G., \& Sachs, U. J. H. (2009). Transfusion-associated graft-versus-host disease. Transfusion Medicine Reviews, 23, 62-71.

Samejima, F. (1997). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. Psychometrika, 62, 471493.

Sato, E., Rabinowitz, S., Gallagher, C., \& Huang, C.-W. (2010). Accommodations for English language learner students: The effect of linguistic modification of math test item sets. Final Report. NCEE 2009-4079. June 2010, English Document (ED) 204. National Center for Education Evaluation and Regional Assistance. Available from: ED Pubs. P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827; Web site: http://ies.ed.gov.libproxy.library.wmich.edu/ncee/

Schlesselman, J. J. (1982). Case-control studies. New York: Oxford University Press.

Silberman, S. (1999). Platelets: Preparations, transfusion, modifications, and substitutes. Archives of Pathology \& Laboratory Medicine, 123, 889-894.

Sinha, S., Mallick, B. K, Kipnis, V., \& Carroll, R. J. (2010). Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. Biometrics, 66, 444-454.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. Journal of Educational Measurement, 42, 375-394.

Sinharay, S., Johnson, M. S., \& Stern, H. S. (2006). Posterior predictive assessment of item response theory models. Applied Psychological Measurement, 30, 298-321.

Steinberg, L., \& Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods, 11, 402-415.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., \& Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology, 21, 128138.

Stone, C. A. (2003). Empirical power and type I error rates for an IRT fit statistic that considers the precision of ability estimates. Educational and Psychological Measurement, 63, 566-583.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stute, W., Manteiga, W. G., \& Quindmil, M. P. (1993). Bootstrap based goodness-of-fittests. Metrika, 40, 243-256.

Swaminathan, H., \& Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50, 349-364.

Swaminathan, H., Hambleton, R. K., Sireci, S. D., Xing, D., \& Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates (Law School Admission Council Computerized Testing Report 98-06). Applied Psychological Measurement, 27(1), 27-51.

Swaminathan, H., \& Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Tang, M.-L., Tang, N.-S., Chan, I. S.-F., \& Chan, B. P.-S. (2002). Sample size determination for establishing equivalence/noninferiority via ratio of two proportions in matched-pair design. Biometrics, 58, 957-963.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Statistics in Medicine, 17, 891-908.

Teresi, J. A., Kleinman, M., \& Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Statistics in Medicine, 19, 1651-1683.

Thompson, N. A. (2009). Ability estimation with item response theory (White Paper). St. Paul, MN: Assessment Systems Corporation.

Toribio, S. G. (2006). Bayesian model checking strategies for dichotomous item response theory models. Doctoral dissertation, Bowling Green State University. Scientific Commons. Retrieved from http://etd.ohiolink.edu/send-pdf.cgi/Toribio\  Sherwin\%20G.pdf?bgsu1150425606
U.S. Department of Health and Human Services. (2007a). Guidance for industry and FDA review staff: Collection of platelets by automated methods. Rockville, MD: Author.
U.S. Department of Health and Human Services. (2007b). 2007 National Blood Collection and Utilization Survey. Rockville, MD: Author.
U.S. Department of Health and Human Services. (2010a). Guidance for industry: Noninferiority clinical trials. Rockville, MD: Author.
U.S. Department of Health and Human Services. (2010b). Guidance for industry and staff: Guidance for the use of Bayesian statistics in medical device clinical trials. Rockville, MD: Author.
van der Linden, W. J., \& Hambleton, R. K. (1997). Handbook of modern item response theory. New York: Springer.

Van Ness, P. H., Holford, T. R., \& Dubin, J. A. (2005, April). Power simulation for categorical data: Using the RANTBL function. Paper presented at SUGI 30, Philadelphia, PA.

Veenstra, D. L., Harris, J., Gibson, R. L., Rosenfeld, M., Burke, W., \& Watts, C. (2007). Pharmacogenomic testing to prevent aminoglycoside-induced hearing loss in cystic fibrosis patients: Potential impact on clinical, patient, and economic outcomes. Genetics in Medicine, 9, 695-704.

Wacholder, S., \& Weinberg, C. R. (1982). Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: Power considerations. Biometrics, 38, 801-812.

Wasson, J. H., Sox, H. C., Neff, R. K., \& Goldman, L. (1985). Clinical prediction rules: Applications and methodological standards. The New England Journal of Medicine, 313, 793-799.

Wiens, B. L. (2002). Choosing and equivalence limit for noninferiority or equivalence studies. Controlled Clinical Trials, 23, 2-14.

Wiens, B. L. (2006). Randomization as a basis for inference in noninferiority trials. Pharmaceutical Statistics, 5, 265-271.

Wright, B. D., \& Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.

Yeh, K. C., \& Kwan, K. C. (1978). A comparison of numerical integrating algorithms by trapezoidal, Lagrange, and spline approximation. Journal of Pharmacokinetics and Biopharmaceutics, 6, 79-98.

Yeh, S.-T. (2002, April). Using trapezoidal rule for the area under a curve calculation. Paper presented at SUGI 27, Orlando, FL.

Yen, W. M. (1981). Using simultaneous results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Zhang, J., \& Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. Psychometrika, 70, 203-212.

Zou, K. H., Gastwirth, J. L., \& McNeil, B. J. (2003). A goodness-of-fit test for a receiver operating characteristic curve from continuous diagnostic test data. Lecture NotesMonograph Series, 43, 60-68.

Zwinderman, A. H. (2001, June). Twisting IRT models for evaluation of treatment related side-effects. Atelier 127. Statistical methods for measurement of subjective variables in health. Presented at the 2001 conference.

Appendix A
Additional Convergence Plots for 2-PL IRT Model


Figure A-1. Trace and G-R Plot for Mean Discrimination $(n=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-2. Trace and G-R Plot for Mean Difficulty $(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-3. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-4. Trace and G-R Plot for SD Discrimination $(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-5. Trace and G-R Plot for SD Difficulty $(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-6. Trace and G-R Plot for AE Predisposition $(\boldsymbol{\theta})(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-7. Trace and G-R Plot for Mean Discrimination $(n=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-8. Trace and G-R Plot for Mean Difficulty $(n=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-9. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathrm{i})}=\mathbf{1}$


Figure A-10. Trace and G-R Plot for SD Discrimination $(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-11. Trace and G-R Plot for SD Difficulty $(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure $A-12$. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(n=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-13. Trace and G-R Plot for Mean Discrimination $(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-14. Trace and G-R Plot for Mean Difficulty $(n=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-15. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-16. Trace and G-R Plot for SD Discrimination $(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-17. Trace and G-R Plot for SD Difficulty $(n=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-18. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure A-19. Trace and G-R Plot for Mean Discrimination $(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-20. Trace and G-R Plot for Mean Difficulty $(n=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-21. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(n=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-22. Trace and G-R Plot for SD Discrimination $(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure A-23. Trace and G-R Plot for SD Difficulty $(n=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$

|  |  |
| :---: | :---: |

Figure A-24. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(n=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}$,

$$
\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}
$$

## Appendix B

Additional Convergence Plots for 2-PL EX IRT Model


Figure B-1. Trace and G-R Plot for Mean Discrimination (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$

| AE1: Blue; AE2: Red; AE3: Green <br> AE4: Cyan; AE5: Orange; AE6: Black <br> AE7: Brown; AE8: Olive; AE9: Gold | AE1: Blue; AE2: Red; AE3: Green <br> AE4: Cyan; AE5: Orange; AE6: Black <br> AE7: Brown; AE8: Olive; AE9: Gold |
| :---: | :---: |
|  |  |

Figure B-2. Trace and G-R Plot for Mean Difficulty (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-3. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-4. Trace and G-R Plot for SD Discrimination (SRS =30): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-5. Trace and G-R Plot for SD Difficulty $($ SRS $=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7 ,} \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-6. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-7. Trace and G-R Plot for Mean Discrimination $(\mathrm{SRS}=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-8. Trace and G-R Plot for Mean Difficulty (SRS = 30): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-9. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-10. Trace and G-R Plot for SD Discrimination (SRS = 30): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-11. Trace and G-R Plot for SD Difficulty (SRS =30): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-12. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-13. Trace and G-R Plot for Mean Discrimination (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-14. Trace and G-R Plot for Mean Difficulty (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-15. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-16. Trace and G-R Plot for SD Discrimination (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-17. Trace and G-R Plot for SD Difficulty (SRS =250): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-18. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure B-19. Trace and G-R Plot for Mean Discrimination $(\mathrm{SRS}=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-20. Trace and G-R Plot for Mean Difficulty (SRS = 250): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-21. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-22. Trace and G-R Plot for SD Discrimination $(\mathrm{SRS}=250)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-23. Trace and G-R Plot for SD Difficulty $($ SRS $=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure B-24. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$

Appendix C
Additional Convergence Plots for 2-PL MEX IRT Model


Figure C-1. Trace and G-R Plot for Mean Discrimination (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-2. Trace and G-R Plot for Mean Difficulty (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-3. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-4. Trace and G-R Plot for SD Discrimination (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-5. Trace and G-R Plot for SD Difficulty $($ SRS $=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7 ,} \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-6. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30)$ : $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-7. Trace and G-R Plot for Mean Discrimination $(\mathrm{SRS}=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-8. Trace and G-R Plot for Mean Difficulty (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-9. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-10. Trace and G-R Plot for SD Discrimination $(S R S=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-11. Trace and G-R Plot for SD Difficulty (SRS $=30$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-12. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=30): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-13. Trace and G-R Plot for Mean Discrimination (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-14. Trace and G-R Plot for Mean Difficulty (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-15. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-16. Trace and G-R Plot for SD Discrimination (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-17. Trace and G-R Plot for SD Difficulty $(\operatorname{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-18. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{7}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{0}$


Figure C-19. Trace and G-R Plot for Mean Discrimination $(\operatorname{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-20. Trace and G-R Plot for Mean Difficulty (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-21. Trace and G-R Plot for Mean AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathrm{i})}=\mathbf{1}$


Figure C-22. Trace and G-R Plot for SD Discrimination (SRS $=250$ ): $\mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-23. Trace and G-R Plot for SD Difficulty $(\operatorname{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}, \boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$


Figure C-24. Trace and G-R Plot for SD AE Predisposition $(\boldsymbol{\theta})(\mathrm{SRS}=250): \mathbf{A}^{(\mathbf{i})}=\mathbf{2}$, $\boldsymbol{\theta}^{(\mathbf{i})}=\mathbf{1}$

Appendix D
2-PL IRT Model: Goodness-of-Fit

Table D-1. 2-PL: Bock's Index for AE Type 2

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 27.5250 | 3.54840 | 27.1146 | .000275637 | 27.9353 | .000193272 |
| 250 | 249.485 | 15.0607 | 247.744 | 0 | 251.227 | 0 |



Figure D-1. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $2(n=30)$


Figure D-2. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $2(n=250)$

Table D-2. 2-PL IRT Model - Bock's Index for AE Type 3

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 25.7739 | 3.40284 | 25.3804 | .000580570 | 26.1673 | .000414428 |
| 250 | 162.588 | 10.8993 | 161.328 | 0 | 163.848 | 0 |



Figure D-3. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $3(n=30)$


Figure D-4. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type 3 ( $n=250$ )

Table D-3. 2-PL IRT Model-Bock's Index for AE Type 4

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 23.3383 | 3.65949 | 22.9152 | .001650801 | 23.7615 | .001155444 |
| 250 | 186.987 | 12.0271 | 185.596 | 0 | 188.377 | 0 |



Figure D-5. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $4(n=30)$


Figure D-6. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $4(n=250)$

Table D-4. 2-PL IRT Model - Bock's Index for AE Type 5

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 23.6638 | 3.25462 | 23.2874 | .001411421 | 24.0401 | .001026908 |
| 250 | 175.035 | 11.1706 | 173.744 | 0 | 176.327 | 0 |



Figure D-7. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $5(n=30)$


Figure D-8. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type 5 ( $n=250$ )

Table D-5. 2-PL IRT Model - Bock's Index for AE Type 6

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 22.0738 | 3.39111 | 21.6817 | .002765245 | 22.4660 | .001993130 |
| 250 | 221.405 | 13.8203 | 219.807 | 0 | 223.003 | 0 |



Figure D-9. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $6(n=30)$


Figure D-10. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $6(n=250)$

Table D-6. 2-PL IRT Model - Bock's Index for AE Type 7

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 30.5827 | 3.45697 | 30.1830 | .000072563 | 30.9825 | .000051090 |
| 250 | 238.304 | 14.1529 | 236.667 | 0 | 239.941 | 0 |



Figure D-11. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $7(n=30)$


Figure D-12. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type 7 ( $n=250$ )

Table D-7. 2-PL IRT Model - Bock's Index for AE Type 8

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 24.0098 | 3.40132 | 23.6165 | .001228466 | 24.4031 | .000880368 |
| 250 | 241.944 | 15.2118 | 240.185 | 0 | 243.703 | 0 |



Figure D-13. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $8(n=30)$


Figure D-14. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $8(n=250)$

Table D-8. 2-PL IRT Model - Bock's Index for AE Type 9

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 22.8545 | 3.86330 | 22.4077 | .002042336 | 23.3012 | .001403258 |
| 250 | 169.275 | 12.3739 | 167.844 | 0 | 170.706 | 0 |



Figure D-15. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type $9(n=30)$


Figure D-16. 2-PL IRT Model - Bayesian PPI and Residual Plot for AE Type 9 ( $n=250$ )

## Appendix E

2-PL IRT Model: AE Predisposition Estimates

Table E-1. AE Predisposition Parameter for $n=30$ Patients

| Patient | Mean (SD) | Median | 95\% CI <br> on Mean | Patient | Mean (SD) | Median | 95\% CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $0.07(0.023)$ | 0.07 | $(0.04,0.10)$ | 16 | $0.16(0.015)$ | 0.16 | $(0.14,0.18)$ |
| 2 | $0.03(0.039)$ | 0.05 | $(-0.02,0.08)$ | 17 | $0.19(0.012)$ | 0.19 | $(0.18,0.21)$ |
| 3 | $0.21(0.011)$ | 0.21 | $(0.20,0.23)$ | 18 | $0.05(0.014)$ | 0.05 | $(0.03,0.07)$ |
| 4 | $0.07(0.006)$ | 0.07 | $(0.06,0.07)$ | 19 | $0.04(0.007)$ | 0.04 | $(0.03,0.05)$ |
| 5 | $0.11(0.077)$ | 0.15 | $(0.02,0.21)$ | 20 | $0.09(0.017)$ | 0.09 | $(0.07,0.11)$ |
| 6 | $-0.04(0.016)$ | -0.05 | $(-0.06,-0.02)$ | 21 | $-0.20(0.014)$ | -0.20 | $(-0.22,-0.18)$ |
| 7 | $0.01(0.013)$ | 0.00 | $(-0.01,0.02)$ | 22 | $0.29(0.016)$ | 0.30 | $(0.27,0.31)$ |
| 8 | $0.07(0.004)$ | 0.07 | $(0.06,0.07)$ | 23 | $0.13(0.018)$ | 0.12 | $(0.10,0.15)$ |
| 9 | $-0.03(0.012)$ | -0.03 | $(-0.04,-0.01)$ | 24 | $-0.02(0.019)$ | -0.02 | $(-0.04,0.01)$ |
| 10 | $0.06(0.013)$ | 0.06 | $(0.05,0.08)$ | 25 | $0.08(0.019)$ | 0.08 | $(0.06,0.11)$ |
| 11 | $-0.01(0.015)$ | -0.02 | $(-0.03,0.01)$ | 26 | $0.02(0.014)$ | 0.01 | $(0.00,0.04)$ |
| 12 | $0.15(0.010)$ | 0.15 | $(0.14,0.16)$ | 27 | $0.03(0.028)$ | 0.02 | $(-0.01,0.06)$ |
| 13 | $0.06(0.026)$ | 0.05 | $(0.03,0.09)$ | 28 | $-0.04(0.095)$ | -0.01 | $(-0.16,0.08)$ |
| 14 | $-0.01(0.013)$ | -0.01 | $(-0.03,0.00)$ | 29 | $-0.16(0.023)$ | -0.17 | $(-0.19,-0.13)$ |
| 15 | $0.01(0.010)$ | 0.02 | $(0.00,0.02)$ | 30 | $-0.10(0.006)$ | -0.10 | $(-0.11,-0.09)$ |

Table E-2. AE Predisposition Parameter for $n=250$ Patients

| Patient | Mean (SD) | Median | 95\% CI <br> on Mean | Patient | Mean (SD) | Median | 95\% CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $-0.02(0.026)$ | -0.02 | $(-0.05,0.01)$ | 126 | $0.10(0.008)$ | 0.11 | $(0.09,0.11)$ |
| 2 | $-0.08(0.026)$ | -0.08 | $(-0.11,-0.04)$ | 127 | $-0.01(0.029)$ | -0.02 | $(-0.05,0.03)$ |
| 3 | $-0.22(0.003)$ | -0.22 | $(-0.22,-0.22)$ | 128 | $-0.05(0.021)$ | -0.05 | $(-0.07,-0.02)$ |
| 4 | $-0.11(0.014)$ | -0.11 | $(-0.13,-0.09)$ | 129 | $-0.11(0.013)$ | -0.11 | $(-0.12,-0.09)$ |
| 5 | $-0.19(0.032)$ | -0.17 | $(-0.23,-0.15)$ | 130 | $-0.07(0.022)$ | -0.06 | $(-0.10,-0.04)$ |
| 6 | $0.19(0.024)$ | 0.18 | $(0.16,0.22)$ | 131 | $-0.08(0.015)$ | -0.08 | $(-0.10,-0.06)$ |
| 7 | $-0.10(0.015)$ | -0.11 | $(-0.12,-0.09)$ | 132 | $-0.11(0.016)$ | -0.11 | $(-0.13,-0.09)$ |
| 8 | $0.01(0.075)$ | -0.01 | $(-0.08,0.10)$ | 133 | $-0.07(0.022)$ | -0.07 | $(-0.09,-0.04)$ |
| 9 | $-0.12(0.019)$ | -0.12 | $(-0.15,-0.10)$ | 134 | $-0.04(0.009)$ | -0.04 | $(-0.06,-0.03)$ |
| 10 | $0.11(0.010)$ | 0.10 | $(0.09,0.12)$ | 135 | $-0.18(0.005)$ | -0.18 | $(-0.19,-0.17)$ |
| 11 | $-0.13(0.006)$ | -0.13 | $(-0.13,-0.12)$ | 136 | $0.27(0.012)$ | 0.27 | $(0.26,0.29)$ |
| 12 | $-0.24(0.014)$ | -0.25 | $(-0.26,-0.22)$ | 137 | $0.02(0.008)$ | 0.02 | $(0.01,0.03)$ |
| 13 | $-0.08(0.021)$ | -0.08 | $(-0.11,-0.05)$ | 138 | $-0.00(0.006)$ | 0.00 | $(-0.01,0.01)$ |


| Patient | Mean (SD) | Median | $\begin{gathered} \text { 95\% CI } \\ \text { on Mean } \end{gathered}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 14 | -0.18(0.004) | -0.18 | $(-0.19,-0.18)$ | 139 | 0.28 (0.017) | 0.29 | (0.26, 0.31$)$ |
| 15 | -0.11 (0.015) | -0.11 | $(-0.12,-0.09)$ | 140 | -0.08 (0.008) | -0.09 | $(-0.09,-0.07)$ |
| 16 | -0.06 (0.023) | -0.06 | $(-0.09,-0.03)$ | 141 | 0.02 (0.018) | 0.02 | $(-0.00,0.04)$ |
| 17 | -0.31 (0.015) | -0.31 | $(-0.33,-0.29)$ | 142 | -0.02 (0.016) | -0.01 | (-0.04, 0.00) |
| 18 | 0.00 (0.012) | 0.00 | $(-0.01,0.02)$ | 143 | 0.01 (0.016) | 0.02 | $(-0.01,0.03)$ |
| 19 | 0.13 (0.012) | 0.13 | (0.12, 0.14) | 144 | -0.05 (0.015) | $-0.05$ | ( $-0.07,-0.03$ ) |
| 20 | 0.03 (0.017) | 0.03 | (0.00, 0.05) | 145 | -0.11 (0.018) | -0.11 | $(-0.13,-0.09)$ |
| 21 | -0.19 (0.012) | -0.19 | $(-0.21,-0.18)$ | 146 | 0.06 (0.100) | 0.03 | $(-0.07,0.18)$ |
| 22 | -0.18 (0.022) | -0.18 | $(-0.21,-0.15)$ | 147 | -0.20 (0.014) | -0.21 | $(-0.22,-0.19)$ |
| 23 | 0.16 (0.015) | 0.16 | (0.14, 0.18) | 148 | -0.11 (0.015) | -0.11 | $(-0.13,-0.09)$ |
| 24 | 0.03 (0.013) | 0.03 | (0.02, 0.05) | 149 | 0.07 (0.018) | 0.07 | $(0.05,0.10)$ |
| 25 | -0.33 (0.009) | -0.33 | $(-0.34,-0.32)$ | 150 | 0.16 (0.021) | 0.16 | $(0.13,0.18)$ |
| 26 | 0.10 (0.013) | 0.10 | (0.09, 0.12) | 151 | -0.00 (0.009) | 0.00 | $(-0.01,0.01)$ |
| 27 | -0.23 (0.017) | $-0.23$ | $(-0.25,-0.21)$ | 152 | -0.18 (0.010) | -0.18 | $(-0.19,-0.16)$ |
| 28 | -0.12 (0.014) | -0.13 | $(-0.14,-0.10)$ | 153 | -0.03 (0.019) | -0.03 | $(-0.05,-0.01)$ |
| 29 | -0.11 (0.024) | -0.12 | $(-0.14,-0.08)$ | 154 | -0.02 (0.011) | -0.02 | $(-0.04,-0.01)$ |
| 30 | -0.12 (0.010) | -0.11 | $(-0.13,-0.10)$ | 155 | $-0.01(0.108)$ | -0.05 | $(-0.14,0.13)$ |
| 31 | -0.01 (0.028) | 0.01 | $(-0.04,0.02)$ | 156 | -0.24 (0.014) | -0.25 | $(-0.26,-0.22)$ |
| 32 | -0.04 (0.036) | -0.03 | $(-0.08,0.01)$ | 157 | 0.00 (0.015) | 0.00 | $(-0.01,0.02)$ |
| 33 | 0.16 (0.006) | 0.16 | $(0.16,0.17)$ | 158 | 0.14 (0.019) | 0.15 | (0.11, 0.16) |
| 34 | -0.06 (0.020) | -0.06 | $(-0.09,-0.04)$ | 159 | 0.12 (0.004) | 0.12 | $(0.12,0.13)$ |
| 35 | -0.06 (0.012) | -0.06 | $(-0.07,-0.04)$ | 160 | -0.08 (0.024) | -0.09 | $(-0.11,-0.05)$ |
| 36 | 0.29 (0.031) | 0.27 | $(0.25,0.32)$ | 161 | $-0.29(0.014)$ | -0.30 | $(-0.31,-0.28)$ |
| 37 | 0.11 (0.012) | 0.11 | $(0.09,0.12)$ | 162 | 0.11 (0.007) | 0.11 | $(0.11,0.12)$ |
| 38 | 0.18 (0.013) | 0.18 | $(0.16,0.19)$ | 163 | $-0.01(0.030)$ | $-0.01$ | $(-0.05,0.02)$ |
| 39 | 0.15 (0.026) | 0.16 | (0.12, 0.19) | 164 | 0.14 (0.012) | 0.13 | (0.12, 0.15) |
| 40 | -0.00 (0.018) | 0.01 | $(-0.02,0.02)$ | 165 | -0.03 (0.023) | -0.03 | $(-0.06,-0.00)$ |
| 41 | -0.17 (0.097) | -0.22 | $(-0.29,-0.05)$ | 166 | 0.24 (0.108) | 0.18 | (0.11, 0.38) |
| 42 | 0.07 (0.019) | 0.07 | $(0.05,0.10)$ | 167 | -0.07 (0.016) | -0.07 | $(-0.09,-0.05)$ |
| 43 | -0.22 (0.019) | $-0.22$ | $(-0.25,-0.20)$ | 168 | 0.19 (0.027) | 0.20 | (0.16, 0.22) |
| 44 | -0.23 (0.013) | $-0.23$ | $(-0.25,-0.21)$ | 169 | -0.05 (0.009) | -0.05 | $(-0.06,-0.04)$ |
| 45 | -0.16 (0.026) | -0.16 | $(-0.19,-0.13)$ | 170 | -0.12 (0.026) | -0.13 | $(-0.15,-0.09)$ |
| 46 | 0.16 (0.017) | 0.16 | (0.14, 0.19) | 171 | -0.09 (0.006) | -0.08 | $(-0.09,-0.08)$ |
| 47 | 0.41 (0.026) | 0.41 | $(0.38,0.44)$ | 172 | -0.15 (0.028) | -0.15 | $(-0.19,-0.12)$ |
| 48 | 0.13 (0.086) | 0.10 | (0.03, 0.24) | 173 | 0.00 (0.006) | 0.00 | $(-0.00,0.01)$ |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 49 | -0.28 (0.015) | -0.28 | $(-0.30,-0.26)$ | 174 | -0.10 (0.017) | -0.10 | $(-0.12,-0.07)$ |
| 50 | -0.05 (0.018) | -0.05 | $(-0.07,-0.02)$ | 175 | 0.04 (0.025) | 0.03 | (0.01, 0.07) |
| 51 | -0.15 (0.025) | -0.15 | $(-0.18,-0.12)$ | 176 | 0.12 (0.015) | 0.12 | (0.10, 0.14) |
| 52 | 0.12 (0.022) | 0.12 | $(0.09,0.15)$ | 177 | -0.26 (0.112) | $-0.31$ | $(-0.40,-0.12)$ |
| 53 | 0.10 (0.017) | 0.10 | (0.08, 0.12) | 178 | -0.19 (0.028) | -0.18 | $(-0.22,-0.15)$ |
| 54 | 0.22 (0.013) | 0.22 | (0.21, 0.24) | 179 | 0.05 (0.081) | 0.03 | $(-0.05,0.15)$ |
| 55 | -0.06 (0.017) | -0.06 | $(-0.08,-0.04)$ | 180 | 0.12 (0.025) | 0.12 | (0.09, 0.15) |
| 56 | 0.15 (0.020) | 0.14 | (0.12, 0.17) | 181 | -0.04 (0.012) | -0.04 | $(-0.06,-0.03)$ |
| 57 | -0.10 (0.015) | -0.10 | $(-0.12,-0.08)$ | 182 | -0.20 (0.116) | $-0.25$ | $(-0.34,-0.05)$ |
| 58 | 0.15 (0.014) | 0.14 | $(0.13,0.16)$ | 183 | -0.06 (0.016) | -0.06 | $(-0.08,-0.04)$ |
| 59 | -0.07 (0.016) | -0.07 | $(-0.09,-0.05)$ | 184 | -0.02 (0.027) | -0.03 | $(-0.06,0.01)$ |
| 60 | -0.06 (0.009) | -0.06 | $(-0.07,-0.05)$ | 185 | 0.08 (0.020) | 0.09 | (0.05, 0.10) |
| 61 | -0.02 (0.018) | -0.02 | $(-0.04,0.00)$ | 186 | -0.17 (0.012) | $-0.16$ | $(-0.18,-0.15)$ |
| 62 | -0.12 (0.012) | $-0.12$ | $(-0.13,-0.10)$ | 187 | 0.18 (0.009) | 0.18 | (0.17, 0.19) |
| 63 | -0.10 (0.009) | -0.10 | $(-0.11,-0.08)$ | 188 | -0.07 (0.009) | -0.07 | $(-0.08,-0.06)$ |
| 64 | -0.17 (0.018) | -0.17 | $(-0.19,-0.14)$ | 189 | -0.12 (0.029) | $-0.13$ | $(-0.16,-0.08)$ |
| 65 | 0.15 (0.018) | 0.15 | (0.12, 0.17) | 190 | 0.13 (0.012) | 0.13 | (0.11, 0.14) |
| 66 | 0.22 (0.024) | 0.22 | $(0.19,0.25)$ | 191 | -0.12 (0.020) | -0.12 | $(-0.14,-0.09)$ |
| 67 | 0.22 (0.010) | 0.23 | (0.21, 0.24) | 192 | -0.11 (0.028) | $-0.10$ | $(-0.14,-0.07)$ |
| 68 | 0.16 (0.016) | 0.17 | $(0.14,0.18)$ | 193 | 0.18 (0.012) | 0.18 | $(0.17,0.20)$ |
| 69 | -0.14 (0.014) | -0.15 | $(-0.16,-0.12)$ | 194 | -0.01 (0.020) | -0.01 | $(-0.04,0.01)$ |
| 70 | -0.27 (0.011) | -0.26 | $(-0.28,-0.26)$ | 195 | -0.29 (0.014) | $-0.30$ | $(-0.31,-0.27)$ |
| 71 | -0.10 (0.008) | $-0.10$ | $(-0.11,-0.09)$ | 196 | -0.18 (0.009) | $-0.19$ | $(-0.20,-0.17)$ |
| 72 | 0.03 (0.016) | 0.03 | (0.01, 0.05) | 197 | -0.06 (0.027) | -0.07 | $(-0.09,-0.03)$ |
| 73 | -0.05 (0.019) | -0.06 | $(-0.07,-0.02)$ | 198 | 0.14 (0.008) | 0.14 | $(0.13,0.15)$ |
| 74 | -0.28 (0.018) | -0.28 | $(-0.31,-0.26)$ | 199 | -0.01 (0.015) | -0.02 | $(-0.03,0.00)$ |
| 75 | -0.00 (0.022) | 0.00 | $(-0.03,0.02)$ | 200 | -0.17 (0.015) | -0.17 | $(-0.18,-0.15)$ |
| 76 | -0.01 (0.007) | -0.02 | $(-0.02,-0.01)$ | 201 | 0.06 (0.006) | 0.06 | $(0.05,0.06)$ |
| 77 | -0.13 (0.018) | $-0.13$ | $(-0.15,-0.11)$ | 202 | 0.12 (0.026) | 0.11 | $(0.09,0.15)$ |
| 78 | 0.17 (0.009) | 0.17 | $(0.16,0.18)$ | 203 | 0.03 (0.017) | 0.03 | (0.01, 0.05) |
| 79 | -0.01 (0.029) | -0.01 | $(-0.04,0.03)$ | 204 | 0.23 (0.024) | 0.23 | (0.20, 0.26) |
| 80 | -0.13 (0.034) | -0.12 | $(-0.18,-0.09)$ | 205 | -0.10 (0.020) | -0.11 | $(-0.13,-0.08)$ |
| 81 | 0.25 (0.017) | 0.24 | $(0.23,0.27)$ | 206 | 0.11 (0.009) | 0.11 | $(0.09,0.12)$ |
| 82 | -0.17 (0.038) | -0.18 | $(-0.22,-0.12)$ | 207 | 0.11 (0.019) | 0.10 | $(0.08,0.13)$ |
| 83 | 0.13 (0.018) | 0.13 | $(0.10,0.15)$ | 208 | 0.33 (0.022) | 0.32 | (0.30, 0.36) |


| Patient | Mean (SD) | Median | $\begin{gathered} \text { 95\% CI } \\ \text { on Mean } \end{gathered}$ | Patient | Mean (SD) | Median | $\begin{gathered} \text { 95\% CI } \\ \text { on Mean } \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 84 | 0.06 (0.065) | 0.05 | $(-0.02,0.14)$ | 209 | -0.04 (0.023) | -0.05 | $(-0.07,-0.01)$ |
| 85 | -0.04 (0.014) | -0.05 | $(-0.06,-0.03)$ | 210 | 0.17 (0.007) | 0.17 | $(0.16,0.17)$ |
| 86 | 0.07 (0.030) | 0.07 | (0.03, 0.11) | 211 | -0.16 (0.023) | -0.15 | $(-0.19,-0.13)$ |
| 87 | 0.11 (0.020) | 0.11 | (0.09, 0.14) | 212 | 0.15 (0.018) | 0.14 | (0.12, 0.17) |
| 88 | 0.16 (0.036) | 0.15 | (0.11, 0.20) | 213 | 0.02 (0.012) | 0.01 | $(-0.00,0.03)$ |
| 89 | 0.19 (0.032) | 0.20 | (0.15, 0.23) | 214 | 0.16 (0.008) | 0.16 | $(0.15,0.17)$ |
| 90 | -0.09 (0.010) | -0.08 | $(-0.10,-0.07)$ | 215 | -0.19 (0.005) | -0.19 | $(-0.20,-0.18)$ |
| 91 | -0.16 (0.016) | -0.17 | $(-0.18,-0.14)$ | 216 | -0.21 (0.011) | -0.21 | $(-0.22,-0.20)$ |
| 92 | 0.27 (0.014) | 0.27 | (0.26, 0.29) | 217 | -0.04 (0.020) | -0.05 | $(-0.07,-0.02)$ |
| 93 | 0.38 (0.018) | 0.38 | (0.36, 0.40) | 218 | -0.08 (0.011) | -0.08 | $(-0.10,-0.07)$ |
| 94 | -0.31 (0.012) | -0.30 | $(-0.32,-0.29)$ | 219 | -0.38 (0.007) | -0.38 | $(-0.39,-0.37)$ |
| 95 | 0.07 (0.109) | 0.04 | $(-0.06,0.21)$ | 220 | 0.14 (0.016) | 0.14 | $(0.13,0.16)$ |
| 96 | -0.22 (0.011) | -0.22 | $(-0.24,-0.21)$ | 221 | -0.09 (0.008) | -0.09 | $(-0.10,-0.08)$ |
| 97 | 0.02 (0.012) | 0.02 | (0.01, 0.04) | 222 | -0.21 (0.020) | -0.20 | $(-0.23,-0.18)$ |
| 98 | -0.22 (0.019) | -0.23 | $(-0.25,-0.20)$ | 223 | -0.09 (0.025) | -0.09 | $(-0.12,-0.06)$ |
| 99 | -0.29 (0.018) | -0.29 | $(-0.32,-0.27)$ | 224 | -0.06 (0.007) | -0.06 | $(-0.07,-0.05)$ |
| 100 | -0.13 (0.017) | $-0.13$ | $(-0.15,-0.11)$ | 225 | 0.16 (0.009) | 0.16 | $(0.15,0.17)$ |
| 101 | 0.03 (0.015) | 0.03 | (0.01, 0.05) | 226 | 0.18 (0.015) | 0.18 | (0.16, 0.20) |
| 102 | 0.06 (0.012) | 0.06 | (0.05, 0.08) | 227 | -0.02 (0.007) | -0.02 | $(-0.03,-0.01)$ |
| 103 | 0.22 (0.073) | 0.19 | $(0.13,0.31)$ | 228 | 0.28 (0.018) | 0.28 | (0.26, 0.30) |
| 104 | -0.13 (0.015) | -0.13 | $(-0.15,-0.11)$ | 229 | 0.04 (0.093) | -0.00 | $(-0.08,0.15)$ |
| 105 | -0.24 (0.009) | -0.24 | $(-0.25,-0.23)$ | 230 | 0.00 (0.020) | 0.00 | $(-0.03,0.03)$ |
| 106 | 0.03 (0.016) | 0.02 | (0.01, 0.05) | 231 | -0.18 (0.018) | -0.18 | $(-0.20,-0.16)$ |
| 107 | -0.13 (0.016) | $-0.13$ | $(-0.15,-0.11)$ | 232 | -0.10 (0.006) | -0.10 | $(-0.10,-0.09)$ |
| 108 | 0.11 (0.007) | 0.11 | (0.10, 0.11) | 233 | -0.22 (0.021) | -0.22 | $(-0.24,-0.19)$ |
| 109 | 0.06 (0.018) | 0.05 | (0.03, 0.08) | 234 | -0.25 (0.010) | -0.24 | $(-0.26,-0.23)$ |
| 110 | 0.18 (0.016) | 0.17 | $(0.15,0.20)$ | 235 | -0.03 (0.010) | -0.03 | $(-0.04,-0.02)$ |
| 111 | 0.11 (0.016) | 0.12 | (0.10, 0.13) | 236 | -0.04 (0.004) | -0.04 | $(-0.04,-0.03)$ |
| 112 | -0.21 (0.018) | -0.21 | $(-0.23,-0.18)$ | 237 | -0.10 (0.007) | -0.10 | $(-0.11,-0.09)$ |
| 113 | 0.11 (0.008) | 0.11 | (0.09, 0.12) | 238 | -0.16 (0.005) | -0.16 | $(-0.16,-0.15)$ |
| 114 | -0.18 (0.037) | -0.19 | $(-0.23,-0.14)$ | 239 | 0.17 (0.011) | 0.17 | $(0.15,0.18)$ |
| 115 | -0.03 (0.020) | -0.02 | $(-0.05,-0.00)$ | 240 | 0.03 (0.022) | 0.03 | (0.00, 0.06) |
| 116 | 0.02 (0.021) | 0.01 | $(-0.00,0.05)$ | 241 | $-0.05(0.010)$ | -0.05 | $(-0.06,-0.04)$ |
| 117 | 0.02 (0.021) | 0.02 | $(-0.01,0.04)$ | 242 | 0.07 (0.010) | 0.07 | $(0.05,0.08)$ |
| 118 | 0.10 (0.021) | 0.09 | (0.07, 0.13) | 243 | -0.08 (0.026) | -0.09 | $(-0.11,-0.05)$ |


| Patient | Mean (SD) | Median | $\mathbf{9 5 \%}$ CI <br> on Mean | Patient | Mean (SD) | Median | $\mathbf{9 5 \%}$ CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 119 | $0.03(0.007)$ | 0.03 | $(0.02,0.04)$ | 244 | $-0.11(0.012)$ | -0.11 | $(-0.12,-0.09)$ |
| 120 | $-0.08(0.024)$ | -0.08 | $(-0.11,-0.05)$ | 245 | $-0.03(0.017)$ | -0.04 | $(-0.05,-0.01)$ |
| 121 | $-0.08(0.014)$ | -0.08 | $(-0.10,-0.06)$ | 246 | $-0.20(0.064)$ | -0.23 | $(-0.28,-0.12)$ |
| 122 | $0.05(0.009)$ | 0.04 | $(0.04,0.06)$ | 247 | $0.17(0.011)$ | 0.16 | $(0.15,0.18)$ |
| 123 | $0.19(0.011)$ | 0.19 | $(0.17,0.20)$ | 248 | $-0.08(0.017)$ | -0.09 | $(-0.10,-0.06)$ |
| 124 | $-0.00(0.015)$ | -0.01 | $(-0.02,0.02)$ | 249 | $0.12(0.015)$ | 0.12 | $(0.10,0.13)$ |
| 125 | $-0.19(0.016)$ | -0.19 | $(-0.21,-0.17)$ | 250 | $-0.05(0.035)$ | -0.05 | $(-0.10,-0.01)$ |

Appendix F
2-PL EX IRT Model: Goodness-of-Fit

Table F-1. 2-PL EX IRT Model-Bock's Index for AE Type 2

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 304.872 | 22.1060 | 302.316 | 0 | 307.428 | 0 |
| 250 | 1832.65 | 430.811 | 1782.84 | 0 | 1882.47 | 0 |



Figure F-1. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 2 ( $n=30$ )


Figure F-2. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 2

$$
(n=250)
$$

Table F-2. 2-PL EX IRT Model-Bock's Index for AE Type 3

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 312.073 | 23.4702 | 309.359 | 0 | 314.787 | 0 |
| 250 | 1852.79 | 439.699 | 1801.95 | 0 | 1903.64 | 0 |



Figure F3. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type $3(n=30)$


Figure F-4. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 3

$$
(n=250)
$$

Table F-3. 2-PL EX IRT Model-Bock's Index for AE Type 4

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 297.410 | 22.7412 | 294.780 | 0 | 300.040 | 0 |
| 250 | 1957.21 | 477.150 | 1902.04 | 0 | 2012.39 | 0 |



Figure F-5. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 4

$$
(n=30)
$$



Figure F-6. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 4 ( $n=250$ )

Table F-4. 2-PL EX IRT Model - Bock's Index for AE Type 5

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 322.988 | 23.6490 | 320.253 | 0 | 325.722 | 0 |
| 250 | 1974.52 | 485.198 | 1918.41 | 0 | 2030.62 | 0 |




Figure F-7. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 5 ( $n=30$ )


Figure F-8. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 5

$$
(n=250)
$$

Table F-5. 2-PL EX IRT Model - Bock's Index for AE Type 6

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 306.324 | 23.2735 | 303.633 | 0 | 309.015 | 0 |
| 250 | 1949.95 | 493.234 | 1892.91 | 0 | 2006.98 | 0 |



Figure F-9. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 6

$$
(n=30)
$$



Figure F-10. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 6
( $n=250$ )

Table F-6. 2-PL EX IRT Model - Bock's Index for AE Type 7

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 318.149 | 23.0358 | 315.485 | 0 | 320.813 | 0 |
| 250 | 1954.26 | 520.646 | 1894.06 | 0 | 2014.47 | 0 |



Figure F-11. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 7 ( $n=30$ )


Figure F-12. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 7 ( $n=250$ )

Table F-7. 2-PL IRT Model-Bock's Index for AE Type 8

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 317.932 | 23.5544 | 315.208 | 0 | 320.656 | 0 |
| 250 | 1966.65 | 467.881 | 1912.54 | 0 | 2020.75 | 0 |



Figure F-13. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 8 ( $n=30$ )


Figure F-14. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 8 ( $n=250$ )

Table F-8. 2-PL IRT Model-Bock's Index for AE Type 9

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 320.999 | 23.8895 | 318.237 | 0 | 323.762 | 0 |
| 250 | 1986.06 | 480.036 | 1930.55 | 0 | 2041.56 | 0 |



Figure F-15. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 9 ( $n=30$ )


Figure F-16. 2-PL EX IRT Model - Bayesian PPI and Residual Plot for AE Type 9 ( $n=250$ )

## Appendix G

2-PL EX IRT Model: AE Predisposition Parameter Estimates

Table G-1. AE Predisposition Parameter for $n=30$ Patients

| Patient | Mean (SD) | Median | 95\% CI on <br> Mean | Patient | Mean (SD) | Median | 95\% CI on <br> Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $-0.55(0.056)$ | -0.55 | $(-0.62,-0.48)$ | 16 | $-0.59(0.043)$ | -0.57 | $(-0.65,-0.54)$ |
| 2 | $-0.59(0.068)$ | -0.55 | $(-0.67,-0.51)$ | 17 | $-0.60(0.033)$ | -0.62 | $(-0.64,-0.56)$ |
| 3 | $-0.55(0.070)$ | -0.53 | $(-0.64,-0.46)$ | 18 | $-0.58(0.060)$ | -0.57 | $(-0.66,-0.51)$ |
| 4 | $-0.56(0.075)$ | -0.52 | $(-0.65,-0.46)$ | 19 | $-0.57(0.074)$ | -0.60 | $(-0.66,-0.48)$ |
| 5 | $-0.54(0.039)$ | -0.54 | $(-0.59,-0.49)$ | 20 | $-0.60(0.049)$ | -0.60 | $(-0.67,-0.54)$ |
| 6 | $-0.56(0.031)$ | -0.56 | $(-0.60,-0.52)$ | 21 | $-0.61(0.077)$ | -0.62 | $(-0.70,-0.51)$ |
| 7 | $-0.55(0.070)$ | -0.55 | $(-0.63,-0.46)$ | 22 | $-0.54(0.063)$ | -0.53 | $(-0.62,-0.46)$ |
| 8 | $-0.63(0.044)$ | -0.62 | $(-0.68,-0.57)$ | 23 | $-0.55(0.028)$ | -0.55 | $(-0.59,-0.52)$ |
| 9 | $-0.60(0.099)$ | -0.60 | $(-0.72,-0.48)$ | 24 | $-0.57(0.032)$ | -0.58 | $(-0.61,-0.53)$ |
| 10 | $-0.56(0.031)$ | -0.54 | $(-0.60,-0.52)$ | 25 | $-0.57(0.072)$ | -0.56 | $(-0.66,-0.48)$ |
| 11 | $-0.55(0.047)$ | -0.56 | $(-0.61,-0.49)$ | 26 | $-0.57(0.037)$ | -0.59 | $(-0.61,-0.52)$ |
| 12 | $-0.59(0.059)$ | -0.60 | $(-0.67,-0.52)$ | 27 | $-0.57(0.079)$ | -0.59 | $(-0.66,-0.47)$ |
| 13 | $-0.56(0.047)$ | -0.56 | $(-0.62,-0.50)$ | 28 | $-0.55(0.066)$ | -0.56 | $(-0.64,-0.47)$ |
| 14 | $-0.54(0.039)$ | -0.57 | $(-0.59,-0.50)$ | 29 | $-0.59(0.032)$ | -0.59 | $(-0.63,-0.55)$ |
| 15 | $-0.51(0.071)$ | -0.48 | $(-0.60,-0.42)$ | 30 | $-0.38(0.358)$ | -0.53 | $(-0.82,0.07)$ |

Table G-2. AE Predisposition Parameter for $n=250$ Patients

| Patient | Mean (SD) | Median | $\mathbf{9 5 \%} \mathbf{C I}$ <br> on Mean | Patient | Mean (SD) | Median | 95\% CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $-0.05(0.240)$ | -0.08 | $(-0.12,0.02)$ | 126 | $-0.06(0.235)$ | -0.10 | $(-0.12,0.01)$ |
| 2 | $-0.04(0.297)$ | -0.08 | $(-0.12,0.04)$ | 127 | $-0.14(0.111)$ | -0.16 | $(-0.17,-0.11)$ |
| 3 | $-0.11(0.114)$ | -0.11 | $(-0.15,-0.08)$ | 128 | $-0.09(0.204)$ | -0.10 | $(-0.15,-0.03)$ |
| 4 | $-0.11(0.131)$ | -0.08 | $(-0.15,-0.08)$ | 129 | $-0.08(0.319)$ | -0.12 | $(-0.17,0.01)$ |
| 5 | $-0.09(0.130)$ | -0.08 | $(-0.12,-0.05)$ | 130 | $-0.14(0.148)$ | -0.13 | $(-0.18,-0.10)$ |
| 6 | $-0.06(0.218)$ | -0.06 | $(-0.12,0.00)$ | 131 | $-0.11(0.140)$ | -0.13 | $(-0.15,-0.07)$ |
| 7 | $-0.12(0.133)$ | -0.13 | $(-0.16,-0.08)$ | 132 | $-0.10(0.307)$ | -0.15 | $(-0.19,-0.01)$ |
| 8 | $-0.06(0.226)$ | -0.07 | $(-0.13,0.00)$ | 133 | $-0.08(0.213)$ | -0.11 | $(-0.14,-0.02)$ |
| 9 | $-0.09(0.212)$ | -0.09 | $(-0.15,-0.03)$ | 134 | $-0.01(0.352)$ | -0.10 | $(-0.11,0.09)$ |
| 10 | $-0.14(0.179)$ | -0.10 | $(-0.19,-0.09)$ | 135 | $-0.10(0.131)$ | -0.10 | $(-0.14,-0.06)$ |
| 11 | $-0.08(0.247)$ | -0.09 | $(-0.15,-0.01)$ | 136 | $-0.10(0.129)$ | -0.10 | $(-0.13,-0.06)$ |
| 12 | $-0.11(0.113)$ | -0.12 | $(-0.14,-0.08)$ | 137 | $-0.03(0.325)$ | -0.07 | $(-0.12,0.06)$ |
| 13 | $-0.09(0.261)$ | -0.12 | $(-0.16,-0.01)$ | 138 | $-0.08(0.128)$ | -0.09 | $(-0.11,-0.04)$ |
| 14 | $-0.03(0.326)$ | -0.09 | $(-0.12,0.07)$ | 139 | $-0.09(0.186)$ | -0.12 | $(-0.14,-0.03)$ |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 15 | -0.09 (0.125) | -0.11 | $(-0.13,-0.06)$ | 140 | -0.07 (0.219) | -0.07 | $(-0.13,-0.01)$ |
| 16 | -0.07 (0.241) | -0.09 | $(-0.13,0.00)$ | 141 | -0.10 (0.133) | -0.09 | $(-0.14,-0.06)$ |
| 17 | -0.08 (0.122) | -0.08 | $(-0.12,-0.05)$ | 142 | -0.10 (0.299) | -0.14 | $(-0.19,-0.02)$ |
| 18 | -0.09 (0.150) | -0.09 | $(-0.13,-0.05)$ | 143 | -0.12 (0.123) | -0.13 | $(-0.16,-0.09)$ |
| 19 | -0.07 (0.223) | -0.10 | (-0.14, -0.01) | 144 | -0.07 (0.142) | -0.07 | $(-0.11,-0.03)$ |
| 20 | -0.07 (0.241) | -0.10 | $(-0.14,-0.00)$ | 145 | -0.09 (0.220) | -0.11 | $(-0.15,-0.02)$ |
| 21 | -0.11 (0.121) | -0.11 | $(-0.15,-0.08)$ | 146 | -0.07 (0.268) | -0.11 | $(-0.15,0.01)$ |
| 22 | -0.12 (0.118) | -0.12 | $(-0.16,-0.09)$ | 147 | -0.12 (0.122) | -0.12 | $(-0.16,-0.09)$ |
| 23 | -0.10 (0.261) | -0.13 | $(-0.17,-0.03)$ | 148 | -0.09 (0.231) | -0.13 | $(-0.15,-0.02)$ |
| 24 | -0.08 (0.245) | -0.10 | $(-0.15,-0.01)$ | 149 | -0.10 (0.115) | -0.10 | $(-0.14,-0.07)$ |
| 25 | -0.05 (0.280) | -0.09 | ( $-0.13,0.02$ ) | 150 | -0.13 (0.123) | -0.13 | $(-0.16,-0.09)$ |
| 26 | -0.09 (0.134) | -0.09 | $(-0.13,-0.05)$ | 151 | -0.05 (0.287) | -0.09 | $(-0.13,0.03)$ |
| 27 | -0.10 (0.143) | -0.09 | $(-0.14,-0.05)$ | 152 | -0.05 (0.222) | -0.07 | $(-0.11,0.01)$ |
| 28 | -0.12 (0.124) | -0.13 | $(-0.15,-0.08)$ | 153 | -0.11 (0.123) | -0.11 | $(-0.15,-0.08)$ |
| 29 | -0.05 (0.311) | -0.09 | $(-0.14,0.04)$ | 154 | -0.03 (0.320) | -0.11 | $(-0.12,0.06)$ |
| 30 | -0.10 (0.233) | -0.14 | $(-0.17,-0.04)$ | 155 | -0.12 (0.127) | -0.12 | $(-0.16,-0.09)$ |
| 31 | -0.11 (0.212) | -0.14 | $(-0.17,-0.05)$ | 156 | -0.07 (0.130) | -0.07 | $(-0.11,-0.04)$ |
| 32 | -0.09 (0.108) | -0.11 | $(-0.12,-0.06)$ | 157 | -0.08 (0.130) | $-0.06$ | $(-0.12,-0.04)$ |
| 33 | -0.11 (0.153) | -0.10 | $(-0.15,-0.06)$ | 158 | -0.03 (0.289) | -0.08 | $(-0.12,0.05)$ |
| 34 | -0.10 (0.134) | -0.12 | $(-0.14,-0.07)$ | 159 | -0.10 (0.233) | $-0.13$ | $(-0.17,-0.04)$ |
| 35 | -0.10 (0.131) | -0.08 | $(-0.14,-0.06)$ | 160 | -0.10 (0.137) | -0.10 | $(-0.14,-0.06)$ |
| 36 | -0.05 (0.348) | -0.12 | $(-0.15,0.05)$ | 161 | -0.10 (0.126) | -0.12 | $(-0.13,-0.06)$ |
| 37 | -0.10 (0.147) | -0.13 | $(-0.14,-0.06)$ | 162 | -0.11 (0.119) | -0.10 | $(-0.14,-0.07)$ |
| 38 | -0.09 (0.123) | -0.12 | $(-0.13,-0.06)$ | 163 | -0.04 (0.248) | $-0.07$ | $(-0.11,0.03)$ |
| 39 | -0.05 (0.296) | -0.11 | $(-0.13,0.04)$ | 164 | -0.15 (0.145) | -0.14 | $(-0.19,-0.11)$ |
| 40 | -0.14 (0.146) | -0.13 | $(-0.18,-0.09)$ | 165 | -0.09 (0.125) | -0.08 | $(-0.12,-0.05)$ |
| 41 | -0.10 (0.228) | -0.13 | $(-0.17,-0.04)$ | 166 | -0.11 (0.140) | -0.11 | $(-0.15,-0.07)$ |
| 42 | -0.10 (0.145) | -0.11 | $(-0.15,-0.06)$ | 167 | -0.13 (0.120) | -0.14 | $(-0.17,-0.10)$ |
| 43 | -0.08(0.132) | -0.08 | $(-0.12,-0.05)$ | 168 | -0.09 (0.222) | -0.12 | $(-0.16,-0.03)$ |
| 44 | -0.08 (0.231) | -0.10 | $(-0.14,-0.01)$ | 169 | -0.11 (0.124) | -0.09 | $(-0.15,-0.08)$ |
| 45 | -0.10 (0.149) | -0.11 | $(-0.14,-0.06)$ | 170 | -0.10 (0.130) | -0.11 | $(-0.14,-0.07)$ |
| 46 | -0.09 (0.231) | -0.10 | $(-0.16,-0.02)$ | 171 | -0.06 (0.147) | -0.09 | $(-0.10,-0.02)$ |
| 47 | -0.11 (0.134) | -0.12 | $(-0.14,-0.07)$ | 172 | -0.09 (0.120) | -0.11 | $(-0.12,-0.05)$ |
| 48 | -0.06 (0.248) | -0.07 | $(-0.13,0.02)$ | 173 | -0.08 (0.307) | -0.12 | $(-0.16,0.01)$ |
| 49 | -0.12 (0.117) | -0.09 | $(-0.15,-0.09)$ | 174 | -0.08 (0.231) | -0.10 | $(-0.15,-0.02)$ |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 50 | -0.11 (0.160) | -0.09 | (-0.16, -0.07) | 175 | -0.08 (0.220) | -0.12 | $(-0.15,-0.02)$ |
| 51 | -0.04 (0.308) | -0.10 | $(-0.13,0.05)$ | 176 | -0.12 (0.236) | -0.14 | $(-0.19,-0.05)$ |
| 52 | -0.07 (0.125) | -0.05 | $(-0.11,-0.04)$ | 177 | -0.08 (0.217) | -0.10 | $(-0.14,-0.02)$ |
| 53 | -0.12 (0.151) | -0.11 | $(-0.16,-0.08)$ | 178 | -0.08 (0.244) | -0.10 | $(-0.15,-0.01)$ |
| 54 | -0.11 (0.215) | -0.14 | $(-0.18,-0.05)$ | 179 | -0.13 (0.126) | -0.11 | $(-0.17,-0.10)$ |
| 55 | -0.06 (0.255) | -0.11 | $(-0.13,0.01)$ | 180 | -0.08 (0.239) | -0.08 | $(-0.15,-0.01)$ |
| 56 | -0.10 (0.231) | -0.13 | $(-0.17,-0.04)$ | 181 | -0.08 (0.241) | -0.10 | $(-0.14,-0.01)$ |
| 57 | -0.07 (0.203) | -0.09 | $(-0.13,-0.01)$ | 182 | -0.06 (0.234) | -0.10 | $(-0.13,0.00)$ |
| 58 | -0.11 (0.144) | -0.12 | $(-0.15,-0.07)$ | 183 | -0.03 (0.281) | -0.07 | $(-0.11,0.05)$ |
| 59 | -0.11 (0.158) | -0.10 | $(-0.16,-0.07)$ | 184 | -0.12 (0.104) | -0.11 | $(-0.15,-0.09)$ |
| 60 | -0.11 (0.248) | -0.12 | $(-0.18,-0.04)$ | 185 | -0.09 (0.149) | $-0.09$ | $(-0.14,-0.05)$ |
| 61 | -0.05 (0.233) | -0.05 | $(-0.12,0.01)$ | 186 | -0.11 (0.132) | -0.14 | $(-0.15,-0.08)$ |
| 62 | -0.11 (0.118) | -0.10 | $(-0.15,-0.08)$ | 187 | -0.11 (0.133) | $-0.13$ | $(-0.15,-0.08)$ |
| 63 | -0.06 (0.279) | -0.11 | $(-0.14,0.02)$ | 188 | -0.12 (0.124) | -0.13 | $(-0.15,-0.08)$ |
| 64 | -0.06 (0.234) | -0.08 | $(-0.13,0.00)$ | 189 | -0.12 (0.155) | $-0.12$ | $(-0.16,-0.07)$ |
| 65 | -0.10 (0.125) | -0.10 | $(-0.14,-0.07)$ | 190 | -0.08 (0.245) | -0.11 | $(-0.15,-0.01)$ |
| 66 | -0.10 (0.228) | -0.13 | $(-0.17,-0.04)$ | 191 | -0.12 (0.125) | -0.13 | $(-0.15,-0.08)$ |
| 67 | -0.10 (0.143) | -0.10 | $(-0.14,-0.06)$ | 192 | -0.10 (0.144) | -0.12 | $(-0.14,-0.06)$ |
| 68 | -0.04 (0.220) | -0.08 | $(-0.10,0.02)$ | 193 | -0.07 (0.210) | -0.10 | $(-0.13,-0.01)$ |
| 69 | -0.14 (0.124) | -0.13 | $(-0.18,-0.11)$ | 194 | -0.09 (0.132) | -0.07 | $(-0.13,-0.05)$ |
| 70 | -0.13 (0.139) | -0.16 | (-0.17, -0.09) | 195 | -0.08 (0.223) | -0.08 | $(-0.14,-0.01)$ |
| 71 | -0.10 (0.133) | -0.09 | $(-0.13,-0.06)$ | 196 | -0.10 (0.123) | -0.12 | $(-0.13,-0.06)$ |
| 72 | -0.05 (0.283) | -0.08 | $(-0.13,0.03)$ | 197 | -0.11 (0.144) | -0.12 | $(-0.15,-0.07)$ |
| 73 | -0.07 (0.237) | -0.09 | (-0.14, -0.00) | 198 | -0.10 (0.119) | -0.10 | $(-0.13,-0.06)$ |
| 74 | -0.10 (0.124) | -0.10 | $(-0.13,-0.06)$ | 199 | -0.09 (0.201) | -0.10 | $(-0.14,-0.03)$ |
| 75 | -0.04 (0.362) | -0.11 | (-0.14, 0.07) | 200 | -0.09 (0.203) | -0.10 | $(-0.14,-0.03)$ |
| 76 | -0.09 (0.128) | -0.11 | (-0.12, -0.05) | 201 | -0.12 (0.113) | -0.14 | $(-0.15,-0.08)$ |
| 77 | -0.02 (0.418) | -0.13 | $(-0.13,0.10)$ | 202 | -0.04 (0.248) | -0.09 | $(-0.11,0.03)$ |
| 78 | -0.01 (0.378) | -0.09 | $(-0.12,0.10)$ | 203 | -0.10 (0.125) | -0.08 | $(-0.13,-0.06)$ |
| 79 | -0.10 (0.135) | -0.12 | $(-0.14,-0.06)$ | 204 | -0.10 (0.220) | -0.10 | $(-0.16,-0.03)$ |
| 80 | -0.11 (0.133) | -0.13 | $(-0.15,-0.07)$ | 205 | -0.11 (0.126) | -0.11 | $(-0.14,-0.07)$ |
| 81 | -0.12 (0.136) | -0.10 | $(-0.16,-0.08)$ | 206 | -0.10 (0.138) | -0.07 | $(-0.14,-0.06)$ |
| 82 | -0.08 (0.223) | -0.09 | $(-0.15,-0.02)$ | 207 | -0.10 (0.121) | -0.08 | $(-0.13,-0.06)$ |
| 83 | -0.09 (0.128) | -0.10 | $(-0.13,-0.06)$ | 208 | -0.10 (0.231) | -0.12 | $(-0.17,-0.04)$ |
| 84 | -0.04 (0.283) | -0.09 | $(-0.12,0.04)$ | 209 | -0.10 (0.128) | -0.07 | $(-0.13,-0.06)$ |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 85 | -0.09 (0.252) | -0.12 | (-0.16, -0.02) | 210 | -0.16 (0.153) | -0.15 | (-0.20, -0.11) |
| 86 | -0.05 (0.291) | -0.08 | $(-0.14,0.03)$ | 211 | -0.00 (0.320) | -0.05 | $(-0.09,0.09)$ |
| 87 | -0.12 (0.144) | -0.08 | $(-0.16,-0.08)$ | 212 | -0.11 (0.150) | -0.11 | $(-0.15,-0.07)$ |
| 88 | -0.10 (0.236) | -0.14 | ( $-0.17,-0.03$ ) | 213 | -0.09 (0.241) | -0.12 | $(-0.15,-0.02)$ |
| 89 | -0.06 (0.211) | -0.06 | (-0.12, -0.00) | 214 | -0.11 (0.210) | -0.15 | $(-0.17,-0.05)$ |
| 90 | -0.10 (0.199) | -0.12 | $(-0.15,-0.04)$ | 215 | -0.12 (0.139) | -0.13 | $(-0.16,-0.08)$ |
| 91 | -0.11 (0.123) | -0.13 | $(-0.15,-0.08)$ | 216 | -0.08 (0.140) | -0.09 | $(-0.12,-0.04)$ |
| 92 | -0.01 (0.295) | -0.05 | $(-0.10,0.07)$ | 217 | -0.11 (0.142) | -0.11 | $(-0.15,-0.07)$ |
| 93 | -0.08 (0.133) | -0.08 | (-0.12, -0.04) | 218 | -0.07 (0.259) | -0.08 | $(-0.14,0.00)$ |
| 94 | -0.08 (0.258) | -0.09 | $(-0.16,-0.01)$ | 219 | -0.09 (0.226) | -0.10 | $(-0.16,-0.03)$ |
| 95 | -0.07 (0.129) | -0.05 | $(-0.11,-0.04)$ | 220 | -0.06 (0.282) | -0.12 | $(-0.14,0.02)$ |
| 96 | -0.09 (0.234) | -0.12 | $(-0.15,-0.02)$ | 221 | -0.09 (0.137) | -0.11 | $(-0.13,-0.05)$ |
| 97 | -0.16 (0.133) | -0.17 | (-0.20, -0.12) | 222 | -0.12 (0.118) | -0.11 | $(-0.15,-0.08)$ |
| 98 | -0.15 (0.136) | -0.16 | $(-0.19,-0.11)$ | 223 | -0.06 (0.219) | -0.07 | $(-0.12,-0.00)$ |
| 99 | -0.08 (0.120) | -0.08 | $(-0.11,-0.04)$ | 224 | -0.11 (0.117) | -0.10 | $(-0.15,-0.08)$ |
| 100 | -0.10 (0.121) | -0.08 | $(-0.13,-0.06)$ | 225 | -0.08 (0.236) | -0.11 | $(-0.15,-0.02)$ |
| 101 | -0.08 (0.229) | -0.11 | $(-0.15,-0.02)$ | 226 | -0.07 (0.209) | -0.10 | $(-0.13,-0.01)$ |
| 102 | -0.09 (0.131) | -0.11 | $(-0.13,-0.05)$ | 227 | -0.11 (0.128) | -0.09 | $(-0.15,-0.07)$ |
| 103 | -0.06 (0.128) | -0.05 | $(-0.10,-0.02)$ | 228 | -0.11 (0.238) | -0.13 | $(-0.18,-0.05)$ |
| 104 | -0.04 (0.319) | -0.08 | $(-0.13,0.05)$ | 229 | -0.04 (0.244) | -0.07 | $(-0.11,0.03)$ |
| 105 | -0.13 (0.147) | -0.12 | $(-0.18,-0.09)$ | 230 | -0.07 (0.242) | -0.11 | $(-0.14,-0.00)$ |
| 106 | -0.12 (0.130) | -0.11 | $(-0.16,-0.09)$ | 231 | -0.05 (0.304) | -0.09 | $(-0.14,0.03)$ |
| 107 | -0.05 (0.299) | -0.12 | $(-0.13,0.04)$ | 232 | -0.10 (0.256) | -0.12 | $(-0.17,-0.03)$ |
| 108 | -0.10 (0.250) | -0.14 | (-0.17, -0.03) | 233 | -0.11 (0.124) | -0.11 | $(-0.15,-0.08)$ |
| 109 | -0.15 (0.124) | -0.17 | $(-0.18,-0.11)$ | 234 | -0.09 (0.131) | -0.07 | $(-0.13,-0.06)$ |
| 110 | -0.07 (0.137) | -0.09 | (-0.11, -0.03) | 235 | -0.09 (0.216) | -0.11 | $(-0.16,-0.03)$ |
| 111 | -0.12 (0.127) | -0.15 | (-0.16, -0.09) | 236 | -0.13 (0.140) | -0.13 | $(-0.17,-0.09)$ |
| 112 | -0.08 (0.253) | -0.13 | $(-0.15,-0.01)$ | 237 | -0.11 (0.124) | -0.13 | $(-0.15,-0.08)$ |
| 113 | -0.10 (0.107) | -0.09 | (-0.14, -0.07) | 238 | -0.10 (0.119) | -0.11 | $(-0.13,-0.06)$ |
| 114 | -0.11 (0.105) | -0.10 | ( $-0.14,-0.08$ ) | 239 | -0.13 (0.147) | -0.12 | $(-0.17,-0.09)$ |
| 115 | -0.09 (0.121) | -0.09 | $(-0.12,-0.05)$ | 240 | -0.08 (0.237) | -0.11 | $(-0.15,-0.01)$ |
| 116 | -0.09 (0.124) | -0.08 | $(-0.12,-0.05)$ | 241 | -0.09 (0.208) | -0.09 | $(-0.15,-0.03)$ |
| 117 | -0.13 (0.123) | -0.15 | $(-0.17,-0.10)$ | 242 | -0.09 (0.144) | -0.09 | $(-0.13,-0.05)$ |
| 118 | -0.07 (0.227) | -0.10 | $(-0.13,-0.00)$ | 243 | -0.11 (0.146) | -0.11 | $(-0.15,-0.07)$ |
| 119 | -0.11 (0.138) | -0.11 | $(-0.15,-0.07)$ | 244 | -0.06 (0.215) | -0.08 | $(-0.12,0.00)$ |


| Patient | Mean (SD) | Median | 95\% CI <br> on Mean | Patient | Mean (SD) | Median | 95\% CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 120 | $-0.07(0.144)$ | -0.07 | $(-0.11,-0.03)$ | 245 | $-0.06(0.294)$ | -0.11 | $(-0.14,0.03)$ |
| 121 | $-0.08(0.216)$ | -0.09 | $(-0.14,-0.01)$ | 246 | $-0.10(0.232)$ | -0.13 | $(-0.16,-0.03)$ |
| 122 | $-0.16(0.102)$ | -0.18 | $(-0.18,-0.13)$ | 247 | $-0.05(0.292)$ | -0.09 | $(-0.13,0.04)$ |
| 123 | $-0.10(0.139)$ | -0.10 | $(-0.14,-0.06)$ | 248 | $-0.14(0.135)$ | -0.16 | $(-0.18,-0.10)$ |
| 124 | $-0.08(0.145)$ | -0.06 | $(-0.12,-0.03)$ | 249 | $0.00(0.349)$ | -0.04 | $(-0.09,0.10)$ |
| 125 | $-0.11(0.135)$ | -0.12 | $(-0.15,-0.07)$ | 250 | $-0.09(0.219)$ | -0.13 | $(-0.15,-0.03)$ |

Appendix H

## 2-PL MEX IRT Model: Goodness-of-Fit

Table H-1. 2-PL MEX IRT Model - Bock's Index for AE Type 2

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 712.68 | 48.7713 | 707.04 | 0 | 718.32 | 0 |
| 250 | 250.043 | 13.4557 | 248.487 | 0 | 251.599 | 0 |



Figure H-1. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 2 ( $n=30$ )


Figure H-2. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 2 ( $n=250$ )

Table H-2. 2-PL MEX IRT Model - Bock's Index for AE Type 3

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :--- | :---: | :---: | :---: | :---: | :---: |
| 30 | 647.43 | 44.3890 | 642.30 | 0 | 652.56 | 0 |
| 250 | 276.252 | 14.1271 | 274.618 | 0 | 277.885 | 0 |



Figure H-3. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 3 ( $n=30$ )


Figure H-4. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 3 ( $n=250$ )

Table H-3. 2-PL MEX IRT Model - Bock's Index for AE Type 4

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 710.43 | 53.5561 | 704.24 | 0 | 716.62 | 0 |
| 250 | 190.655 | 12.2736 | 189.235 | 0 | 192.074 | 0 |



Figure H-5. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 4 ( $n=30$ )


Figure H-6. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 4 ( $n=250$ )

Table H-4. 2-PL MEX IRT Model - Bock's Index for AE Type 5

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 843.64 | 61.9682 | 836.47 | 0 | 850.81 | 0 |
| 250 | 321.880 | 15.3734 | 320.102 | 0 | 323.658 | 0 |



Figure H-7. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 5 ( $n=30$ )


Figure H-8. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 5 ( $n=250$ )

Table H-5. 2-PL MEX IRT Model - Bock's Index for AE Type 6

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 867.10 | 64.2757 | 859.67 | 0 | 874.54 | 0 |
| 250 | 247.677 | 13.8742 | 246.073 | 0 | 249.282 | 0 |



Theta


Figure H-9. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 6

$$
(n=30)
$$



Figure H-10. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 6 ( $n=250$ )

Table H-6. 2-PL MEX IRT Model - Bock's Index for AE Type 7

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 1095.01 | 78.0469 | 1085.98 | 0 | 1104.03 | 0 |
| 250 | 282.358 | 14.4640 | 280.686 | 0 | 284.031 | 0 |



Figure H-11. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 7 ( $n=30$ )


Figure H-12. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 7

$$
(n=250)
$$

Table H-7. 2-PL MEX IRT Model - Bock's Index for AE Type 8

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 848.69 | 62.8318 | 841.43 | 0 | 855.96 | 0 |
| 250 | 235.861 | 12.9578 | 234.363 | 0 | 237.360 | 0 |



Figure H-13. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 8 ( $n=30$ )


Figure H-14. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 8

$$
(n=250)
$$

Table H-8. 2-PL MEX IRT Model - Bock's Index for AE Type 9

| $\boldsymbol{n}$ | bock | std | LCL | pval_low | UCL | pval_high |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 30 | 823.69 | 64.3222 | 816.25 | 0 | 831.13 | 0 |
| 250 | 258.818 | 14.6508 | 257.124 | 0 | 260.512 | 0 |



Figure H-15. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 9 ( $n=30$ )


Figure H-16. 2-PL MEX IRT Model - Bayesian PPI and Residual Plot for AE Type 9 ( $n=250$ )

## Appendix I

2-PL MEX IRT Model: AE Predisposition Parameter Estimates

Table I-1. AE Predisposition Parameter for $n=30$ Patients

| Patient | Mean (SD) | Median | $\mathbf{9 5 \%}$ CI <br> on Mean | Patient | Mean (SD) | Median | $\mathbf{9 5 \%}$ CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $-0.57(0.324)$ | -0.71 | $(-0.98,-0.17)$ | 16 | $-0.59(0.371)$ | -0.62 | $(-1.05,-0.12)$ |
| 2 | $-0.73(0.442)$ | -0.79 | $(-1.28,-0.18)$ | 17 | $-0.61(0.354)$ | -0.78 | $(-1.05,-0.17)$ |
| 3 | $-0.53(0.334)$ | -0.55 | $(-0.94,-0.11)$ | 18 | $-0.56(0.331)$ | -0.64 | $(-0.97,-0.15)$ |
| 4 | $-0.71(0.417)$ | -0.83 | $(-1.22,-0.19)$ | 19 | $-0.55(0.332)$ | -0.62 | $(-0.97,-0.14)$ |
| 5 | $-0.50(0.487)$ | -0.68 | $(-1.10,0.10)$ | 20 | $-0.58(0.350)$ | -0.68 | $(-1.02,-0.15)$ |
| 6 | $-0.55(0.354)$ | -0.70 | $(-0.99,-0.11)$ | 21 | $-0.68(0.451)$ | -0.81 | $(-1.24,-0.12)$ |
| 7 | $-0.68(0.394)$ | -0.82 | $(-1.17,-0.19)$ | 22 | $-0.69(0.419)$ | -0.91 | $(-1.21,-0.17)$ |
| 8 | $-0.69(0.432)$ | -0.79 | $(-1.23,-0.16)$ | 23 | $-0.78(0.448)$ | -0.88 | $(-1.34,-0.22)$ |
| 9 | $-0.70(0.429)$ | -0.93 | $(-1.23,-0.17)$ | 24 | $-0.59(0.361)$ | -0.67 | $(-1.04,-0.14)$ |
| 10 | $-0.79(0.446)$ | -0.96 | $(-1.35,-0.24)$ | 25 | $-0.66(0.437)$ | -0.73 | $(-1.21,-0.12)$ |
| 11 | $-0.57(0.368)$ | -0.70 | $(-1.03,-0.12)$ | 26 | $-0.64(0.403)$ | -0.75 | $(-1.14,-0.14)$ |
| 12 | $-0.75(0.442)$ | -0.89 | $(-1.30,-0.20)$ | 27 | $-0.67(0.409)$ | -0.74 | $(-1.18,-0.16)$ |
| 13 | $-0.68(0.414)$ | -0.75 | $(-1.20,-0.17)$ | 28 | $-0.72(0.409)$ | -0.87 | $(-1.23,-0.21)$ |
| 14 | $-0.50(0.477)$ | -0.64 | $(-1.09,0.09)$ | 29 | $-0.73(0.439)$ | -0.83 | $(-1.28,-0.19)$ |
| 15 | $-0.54(0.319)$ | -0.68 | $(-0.94,-0.14)$ | 30 | $-0.65(0.375)$ | -0.80 | $(-1.12,-0.19)$ |

Table I-2. AE Predisposition Parameter for $n=250$ Patients

| Patient | Mean (SD) | Median | 95\% CI <br> on Mean | Patient | Mean (SD) | Median | 95\% CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $-0.32(0.449)$ | -0.33 | $(-0.87,0.24)$ | 126 | $-0.19(0.452)$ | -0.09 | $(-0.75,0.37)$ |
| 2 | $0.07(0.531)$ | -0.00 | $(-0.59,0.73)$ | 127 | $0.12(0.312)$ | 0.22 | $(-0.27,0.51)$ |
| 3 | $-0.21(0.647)$ | -0.05 | $(-1.01,0.60)$ | 128 | $-0.11(0.265)$ | -0.08 | $(-0.44,0.22)$ |
| 4 | $0.08(0.418)$ | -0.16 | $(-0.44,0.60)$ | 129 | $-0.06(0.547)$ | -0.11 | $(-0.74,0.62)$ |
| 5 | $-0.16(0.224)$ | -0.27 | $(-0.43,0.12)$ | 130 | $-0.09(0.491)$ | -0.05 | $(-0.70,0.52)$ |
| 6 | $-0.35(0.110)$ | -0.37 | $(-0.49,-0.22)$ | 131 | $0.26(0.429)$ | 0.04 | $(-0.27,0.80)$ |
| 7 | $0.25(0.532)$ | 0.30 | $(-0.42,0.91)$ | 132 | $0.17(0.442)$ | 0.27 | $(-0.38,0.72)$ |
| 8 | $-0.12(0.229)$ | -0.05 | $(-0.40,0.17)$ | 133 | $-0.16(0.323)$ | -0.13 | $(-0.56,0.24)$ |
| 9 | $-0.24(0.445)$ | -0.35 | $(-0.79,0.31)$ | 134 | $0.29(0.310)$ | 0.34 | $(-0.10,0.67)$ |
| 10 | $-0.01(0.381)$ | 0.06 | $(-0.48,0.46)$ | 135 | $-0.15(0.302)$ | -0.16 | $(-0.52,0.23)$ |
| 11 | $-0.04(0.275)$ | 0.00 | $(-0.38,0.30)$ | 136 | $0.16(0.432)$ | 0.12 | $(-0.38,0.69)$ |
| 12 | $0.01(0.258)$ | 0.10 | $(-0.31,0.33)$ | 137 | $0.09(0.049)$ | 0.07 | $(0.03,0.15)$ |
| 13 | $-0.11(0.121)$ | -0.13 | $(-0.26,0.04)$ | 138 | $0.14(0.422)$ | -0.03 | $(-0.38,0.67)$ |
| 14 | $-0.14(0.344)$ | -0.23 | $(-0.57,0.28)$ | 139 | $0.09(0.353)$ | 0.07 | $(-0.35,0.52)$ |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 15 | 0.04 (0.306) | 0.01 | (-0.34, 0.42) | 140 | -0.07 (0.531) | -0.19 | $(-0.73,0.59)$ |
| 16 | 0.15 (0.243) | 0.07 | $(-0.15,0.45)$ | 141 | 0.24 (0.545) | 0.08 | $(-0.43,0.92)$ |
| 17 | 0.10 (0.192) | 0.10 | $(-0.14,0.34)$ | 142 | -0.22 (0.372) | 0.01 | $(-0.68,0.24)$ |
| 18 | -0.12 (0.420) | 0.08 | $(-0.64,0.41)$ | 143 | 0.23 (0.366) | 0.36 | $(-0.23,0.68)$ |
| 19 | 0.05 (0.361) | -0.09 | $(-0.40,0.49)$ | 144 | -0.01 (0.209) | -0.03 | $(-0.27,0.25)$ |
| 20 | -0.01 (0.235) | 0.02 | $(-0.30,0.28)$ | 145 | -0.01 (0.316) | -0.18 | $(-0.40,0.39)$ |
| 21 | -0.01 (0.206) | -0.00 | $(-0.27,0.25)$ | 146 | -0.37 (0.424) | -0.48 | $(-0.90,0.15)$ |
| 22 | -0.02 (0.410) | 0.07 | $(-0.53,0.49)$ | 147 | 0.09 (0.460) | 0.21 | $(-0.49,0.66)$ |
| 23 | -0.24 (0.520) | -0.20 | $(-0.89,0.41)$ | 148 | 0.14 (0.247) | 0.04 | $(-0.16,0.45)$ |
| 24 | -0.07 (0.274) | 0.02 | $(-0.41,0.27)$ | 149 | 0.09 (0.122) | 0.08 | $(-0.06,0.25)$ |
| 25 | 0.14 (0.439) | 0.31 | $(-0.41,0.68)$ | 150 | -0.09 (0.386) | -0.23 | $(-0.56,0.39)$ |
| 26 | -0.18 (0.355) | -0.05 | $(-0.62,0.26)$ | 151 | 0.18 (0.197) | 0.20 | $(-0.07,0.42)$ |
| 27 | 0.06 (0.215) | 0.09 | $(-0.21,0.32)$ | 152 | 0.08 (0.176) | 0.07 | $(-0.14,0.29)$ |
| 28 | 0.02 (0.297) | 0.06 | $(-0.35,0.38)$ | 153 | 0.03 (0.297) | -0.04 | $(-0.34,0.40)$ |
| 29 | -0.17 (0.474) | -0.32 | $(-0.76,0.42)$ | 154 | -0.09 (0.197) | -0.19 | $(-0.33,0.16)$ |
| 30 | -0.04 (0.438) | -0.11 | $(-0.58,0.51)$ | 155 | -0.20 (0.125) | -0.18 | $(-0.35,-0.04)$ |
| 31 | -0.04 (0.450) | -0.33 | $(-0.60,0.52)$ | 156 | -0.24 (0.404) | -0.27 | $(-0.74,0.26)$ |
| 32 | 0.19 (0.181) | 0.22 | $(-0.03,0.42)$ | 157 | 0.02 (0.570) | -0.01 | $(-0.69,0.73)$ |
| 33 | 0.00 (0.273) | 0.01 | $(-0.34,0.34)$ | 158 | -0.05 (0.427) | 0.01 | $(-0.58,0.48)$ |
| 34 | -0.21 (0.387) | -0.28 | $(-0.69,0.27)$ | 159 | 0.04 (0.522) | -0.07 | $(-0.61,0.69)$ |
| 35 | -0.07 (0.574) | 0.13 | $(-0.78,0.64)$ | 160 | 0.09 (0.436) | -0.02 | $(-0.45,0.63)$ |
| 36 | -0.12 (0.231) | -0.17 | $(-0.41,0.16)$ | 161 | -0.08 (0.284) | -0.18 | $(-0.43,0.27)$ |
| 37 | -0.09 (0.182) | -0.05 | $(-0.32,0.14)$ | 162 | 0.09 (0.169) | 0.09 | $(-0.12,0.30)$ |
| 38 | -0.27 (0.244) | -0.21 | $(-0.57,0.04)$ | $163$ | -0.04 (0.526) | $0.09$ | $(-0.69,0.61)$ |
| 39 | -0.24 (0.303) | -0.31 | $(-0.62,0.13)$ | 164 | 0.11 (0.545) | 0.11 | (-0.57, 0.78) |
| 40 | -0.05 (0.339) | -0.05 | $(-0.47,0.37)$ | 165 | 0.08 (0.367) | 0.13 | $(-0.37,0.54)$ |
| 41 | 0.03 (0.519) | -0.16 | $(-0.62,0.67)$ | 166 | -0.02 (0.431) | 0.07 | $(-0.55,0.52)$ |
| 42 | 0.01 (0.278) | -0.07 | $(-0.34,0.35)$ | 167 | -0.15 (0.441) | -0.11 | $(-0.70,0.39)$ |
| 43 | 0.11 (0.231) | 0.16 | $(-0.17,0.40)$ | 168 | 0.17 (0.327) | 0.08 | $(-0.24,0.57)$ |
| 44 | 0.03 (0.474) | 0.01 | $(-0.55,0.62)$ | 169 | 0.08 (0.452) | 0.09 | $(-0.48,0.65)$ |
| 45 | -0.14 (0.218) | -0.02 | $(-0.41,0.13)$ | 170 | $-0.04(0.386)$ | -0.13 | $(-0.52,0.44)$ |
| 46 | -0.11 (0.221) | -0.01 | $(-0.38,0.16)$ | 171 | -0.08 (0.276) | 0.01 | $(-0.43,0.26)$ |
| 47 | 0.03 (0.627) | -0.03 | $(-0.75,0.81)$ | 172 | 0.35 (0.256) | 0.35 | $(0.04,0.67)$ |
| 48 | -0.11 (0.316) | 0.02 | $(-0.50,0.28)$ | 173 | 0.02 (0.370) | 0.01 | $(-0.44,0.48)$ |
| 49 | -0.06 (0.533) | -0.01 | $(-0.72,0.60)$ | 174 | 0.03 (0.375) | 0.11 | (-0.43, 0.50) |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 50 | 0.02 (0.402) | -0.10 | $(-0.48,0.52)$ | 175 | 0.14 (0.338) | 0.26 | $(-0.28,0.56)$ |
| 51 | 0.31 (0.606) | 0.53 | $(-0.44,1.06)$ | 176 | 0.15 (0.310) | 0.17 | $(-0.23,0.54)$ |
| 52 | -0.06 (0.282) | 0.00 | $(-0.41,0.29)$ | 177 | -0.19 (0.347) | -0.21 | $(-0.62,0.24)$ |
| 53 | 0.18 (0.259) | 0.20 | $(-0.14,0.50)$ | 178 | 0.01 (0.299) | -0.14 | $(-0.36,0.38)$ |
| 54 | 0.06 (0.198) | -0.01 | $(-0.19,0.30)$ | 179 | -0.06 (0.373) | 0.04 | $(-0.52,0.40)$ |
| 55 | -0.06 (0.170) | -0.02 | $(-0.27,0.15)$ | 180 | 0.15 (0.384) | 0.13 | $(-0.33,0.62)$ |
| 56 | -0.02 (0.419) | 0.23 | ( $-0.54,0.50$ ) | 181 | -0.23 (0.268) | -0.30 | $(-0.56,0.10)$ |
| 57 | 0.10 (0.434) | 0.17 | ( $-0.44,0.64$ ) | 182 | -0.32 (0.248) | -0.37 | $(-0.63,-0.01)$ |
| 58 | -0.00 (0.685) | 0.11 | $(-0.85,0.85)$ | 183 | 0.01 (0.412) | -0.13 | $(-0.50,0.52)$ |
| 59 | -0.10 (0.378) | -0.14 | $(-0.57,0.37)$ | 184 | 0.23 (0.074) | 0.19 | $(0.14,0.32)$ |
| 60 | -0.01 (0.345) | -0.12 | $(-0.44,0.42)$ | 185 | 0.11 (0.172) | 0.17 | $(-0.10,0.32)$ |
| 61 | -0.36 (0.599) | -0.43 | $(-1.11,0.38)$ | 186 | -0.21 (0.467) | -0.32 | $(-0.79,0.36)$ |
| 62 | -0.07 (0.329) | -0.23 | $(-0.48,0.33)$ | 187 | -0.01 (0.272) | 0.07 | $(-0.35,0.33)$ |
| 63 | 0.15 (0.458) | 0.22 | $(-0.42,0.72)$ | 188 | 0.03 (0.376) | -0.02 | $(-0.43,0.50)$ |
| 64 | 0.20 (0.292) | 0.15 | $(-0.17,0.56)$ | 189 | -0.14 (0.406) | 0.09 | $(-0.65,0.36)$ |
| 65 | -0.09 (0.325) | -0.06 | $(-0.49,0.32)$ | 190 | -0.26 (0.270) | -0.27 | $(-0.59,0.08)$ |
| 66 | -0.15 (0.298) | -0.17 | $(-0.52,0.22)$ | 191 | 0.09 (0.273) | 0.10 | $(-0.25,0.42)$ |
| 67 | -0.11 (0.329) | -0.08 | $(-0.52,0.30)$ | 192 | -0.31 (0.398) | -0.27 | $(-0.80,0.19)$ |
| 68 | 0.25 (0.234) | 0.30 | $(-0.04,0.54)$ | 193 | -0.22 (0.405) | -0.06 | $(-0.73,0.28)$ |
| 69 | 0.19 (0.358) | 0.31 | $(-0.25,0.64)$ | 194 | -0.14 (0.524) | $-0.03$ | $(-0.79,0.52)$ |
| 70 | -0.18 (0.251) | -0.12 | $(-0.49,0.13)$ | 195 | -0.14 (0.363) | -0.16 | $(-0.59,0.31)$ |
| 71 | -0.03 (0.272) | 0.16 | $(-0.36,0.31)$ | 196 | 0.02 (0.305) | 0.00 | $(-0.36,0.40)$ |
| 72 | -0.09 (0.361) | -0.01 | $(-0.54,0.36)$ | 197 | -0.03 (0.223) | -0.15 | $(-0.31,0.25)$ |
| 73 | -0.19 (0.186) | -0.18 | $(-0.43,0.04)$ | 198 | 0.40 (0.395) | $0.44$ | $(-0.09,0.89)$ |
| 74 | -0.04 (0.315) | 0.12 | $(-0.43,0.35)$ | 199 | 0.01 (0.206) | -0.06 | $(-0.25,0.26)$ |
| 75 | 0.20 (0.253) | 0.31 | $(-0.12,0.51)$ | 200 | 0.07 (0.322) | 0.13 | $(-0.33,0.47)$ |
| 76 | 0.06 (0.508) | -0.14 | $(-0.57,0.69)$ | 201 | -0.05 (0.173) | 0.00 | $(-0.27,0.16)$ |
| 77 | -0.18 (0.467) | -0.09 | $(-0.76,0.40)$ | 202 | -0.38 (0.385) | -0.47 | $(-0.86,0.09)$ |
| 78 | -0.10 (0.469) | -0.05 | $(-0.68,0.48)$ | 203 | 0.08 (0.549) | -0.04 | $(-0.60,0.76)$ |
| 79 | -0.03 (0.212) | -0.05 | $(-0.29,0.24)$ | 204 | -0.12 (0.301) | -0.01 | $(-0.49,0.26)$ |
| 80 | -0.17 (0.330) | -0.15 | $(-0.58,0.24)$ | 205 | -0.07 (0.255) | -0.14 | $(-0.39,0.24)$ |
| 81 | 0.13 (0.357) | 0.09 | $(-0.31,0.57)$ | 206 | -0.16 (0.183) | -0.17 | $(-0.38,0.07)$ |
| 82 | 0.23 (0.403) | 0.39 | $(-0.27,0.73)$ | 207 | -0.21 (0.393) | -0.19 | $(-0.69,0.28)$ |
| 83 | 0.01 (0.297) | -0.04 | $(-0.35,0.38)$ | 208 | -0.06 (0.322) | 0.13 | $(-0.46,0.34)$ |
| 84 | -0.14 (0.545) | -0.32 | $(-0.82,0.54)$ | 209 | -0.10 (0.469) | 0.03 | $(-0.68,0.49)$ |


| Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Patient | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 85 | 0.03 (0.367) | 0.15 | $(-0.43,0.48)$ | 210 | -0.26 (0.266) | -0.28 | $(-0.59,0.07)$ |
| 86 | -0.32 (0.516) | -0.10 | $(-0.96,0.32)$ | 211 | 0.07 (0.338) | 0.08 | $(-0.35,0.49)$ |
| 87 | -0.12 (0.294) | -0.23 | $(-0.48,0.25)$ | 212 | -0.24 (0.414) | -0.16 | $(-0.75,0.28)$ |
| 88 | -0.20 (0.237) | -0.22 | $(-0.49,0.10)$ | 213 | 0.01 (0.331) | -0.01 | $(-0.40,0.42)$ |
| 89 | -0.04 (0.218) | -0.00 | $(-0.31,0.23)$ | 214 | -0.00 (0.393) | -0.12 | $(-0.49,0.49)$ |
| 90 | 0.26 (0.324) | 0.46 | $(-0.14,0.66)$ | 215 | -0.01 (0.433) | -0.15 | $(-0.55,0.53)$ |
| 91 | 0.07 (0.242) | 0.09 | $(-0.23,0.37)$ | 216 | -0.00 (0.190) | -0.05 | $(-0.24,0.23)$ |
| 92 | 0.04 (0.432) | 0.20 | $(-0.50,0.57)$ | 217 | -0.02 (0.196) | -0.07 | $(-0.26,0.22)$ |
| 93 | -0.03 (0.492) | -0.02 | $(-0.64,0.59)$ | 218 | -0.27 (0.300) | -0.15 | $(-0.65,0.10)$ |
| 94 | 0.13 (0.177) | 0.17 | $(-0.09,0.35)$ | 219 | -0.16 (0.360) | -0.27 | $(-0.60,0.29)$ |
| 95 | -0.16 (0.445) | 0.00 | $(-0.71,0.39)$ | 220 | 0.09 (0.122) | 0.05 | $(-0.06,0.24)$ |
| 96 | 0.38 (0.537) | 0.17 | $(-0.28,1.05)$ | 221 | -0.07 (0.428) | 0.00 | $(-0.60,0.46)$ |
| 97 | 0.02 (0.392) | -0.07 | $(-0.47,0.50)$ | 222 | -0.16 (0.416) | -0.16 | $(-0.68,0.35)$ |
| 98 | 0.24 (0.510) | 0.56 | $(-0.40,0.87)$ | 223 | -0.19 (0.382) | -0.12 | $(-0.67,0.28)$ |
| 99 | 0.24 (0.244) | 0.22 | $(-0.06,0.54)$ | 224 | 0.04 (0.379) | -0.04 | $(-0.43,0.51)$ |
| 100 | -0.01 (0.524) | 0.20 | $(-0.66,0.64)$ | 225 | 0.01 (0.298) | 0.06 | $(-0.36,0.38)$ |
| 101 | -0.14 (0.315) | -0.15 | $(-0.53,0.25)$ | 226 | -0.13 (0.456) | -0.20 | $(-0.69,0.44)$ |
| 102 | 0.03 (0.507) | 0.24 | $(-0.60,0.66)$ | 227 | 0.18 (0.337) | 0.32 | $(-0.24,0.60)$ |
| 103 | -0.10 (0.511) | 0.22 | $(-0.74,0.53)$ | 228 | 0.14 (0.354) | 0.23 | $(-0.30,0.58)$ |
| 104 | 0.05 (0.342) | 0.13 | $(-0.37,0.47)$ | 229 | 0.07 (0.315) | -0.10 | $(-0.32,0.46)$ |
| 105 | -0.45 (0.291) | -0.53 | $(-0.81,-0.09)$ | 230 | 0.11 (0.290) | 0.14 | $(-0.25,0.47)$ |
| 106 | -0.17 (0.283) | -0.26 | $(-0.52,0.19)$ | 231 | -0.00 (0.504) | 0.16 | $(-0.63,0.62)$ |
| 107 | -0.17 (0.266) | -0.27 | $(-0.50,0.16)$ | 232 | -0.03 (0.689) | 0.24 | $(-0.88,0.83)$ |
| 108 | -0.01 (0.317) | -0.14 | $(-0.41,0.38)$ | 233 | 0.02 (0.368) | -0.04 | $(-0.43,0.48)$ |
| 109 | 0.03 (0.264) | -0.08 | $(-0.30,0.36)$ | 234 | 0.01 (0.208) | 0.02 | $(-0.25,0.27)$ |
| 110 | 0.07 (0.465) | -0.08 | $(-0.51,0.64)$ | 235 | -0.06 (0.309) | -0.23 | $(-0.45,0.32)$ |
| 111 | -0.11 (0.270) | -0.10 | $(-0.44,0.23)$ | 236 | -0.02 (0.270) | -0.03 | $(-0.36,0.31)$ |
| 112 | 0.15 (0.232) | 0.26 | $(-0.13,0.44)$ | 237 | 0.08 (0.577) | -0.00 | $(-0.64,0.80)$ |
| 113 | -0.07 (0.430) | -0.06 | $(-0.61,0.46)$ | 238 | 0.16 (0.207) | 0.18 | $(-0.10,0.42)$ |
| 114 | -0.08 (0.412) | -0.11 | $(-0.59,0.43)$ | 239 | -0.02 (0.331) | -0.03 | $(-0.43,0.39)$ |
| 115 | -0.12 (0.388) | -0.27 | $(-0.60,0.36)$ | 240 | 0.11 (0.312) | 0.05 | $(-0.28,0.49)$ |
| 116 | 0.02 (0.395) | 0.01 | $(-0.47,0.51)$ | 241 | 0.10 (0.218) | 0.02 | $(-0.17,0.37)$ |
| 117 | 0.08 (0.204) | 0.10 | $(-0.17,0.34)$ | 242 | -0.03 (0.480) | -0.28 | $(-0.63,0.57)$ |
| 118 | 0.08 (0.417) | 0.06 | $(-0.43,0.60)$ | 243 | 0.11 (0.308) | -0.06 | $(-0.27,0.50)$ |
| 119 | 0.08 (0.281) | 0.10 | $(-0.27,0.43)$ | 244 | -0.23 (0.153) | -0.17 | $(-0.42,-0.04)$ |


| Patient | Mean (SD) | Median | 95\% CI <br> on Mean | Patient | Mean (SD) | Median | 95\% CI <br> on Mean |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 120 | $-0.09(0.413)$ | 0.01 | $(-0.61,0.42)$ | 245 | $-0.12(0.309)$ | -0.21 | $(-0.50,0.27)$ |
| 121 | $-0.12(0.265)$ | -0.15 | $(-0.45,0.20)$ | 246 | $-0.01(0.228)$ | -0.14 | $(-0.30,0.27)$ |
| 122 | $-0.02(0.329)$ | -0.14 | $(-0.43,0.39)$ | 247 | $-0.03(0.333)$ | -0.02 | $(-0.44,0.38)$ |
| 123 | $-0.18(0.468)$ | -0.25 | $(-0.77,0.40)$ | 248 | $0.07(0.223)$ | 0.05 | $(-0.21,0.34)$ |
| 124 | $0.19(0.420)$ | 0.08 | $(-0.33,0.71)$ | 249 | $-0.39(0.261)$ | -0.44 | $(-0.71,-0.07)$ |
| 125 | $0.11(0.294)$ | 0.09 | $(-0.25,0.48)$ | 250 | $0.19(0.368)$ | 0.13 | $(-0.27,0.65)$ |

## Appendix J

2-PL EX IRT Model: Final Estimates of IRT Parameters

Table J-1. Final AE Predisposition Parameter for $n=30$ Patients

| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 1 | -0.55 (0.056) | -0.55 | ( $-0.62,-0.48$ ) | -0.60 (0.042) | -0.58 | $(-0.65,-0.54)$ |
| 2 | -0.59 (0.068) | -0.55 | (-0.67, -0.51) | -0.56 (0.066) | -0.57 | $(-0.64,-0.48)$ |
| 3 | -0.55 (0.070) | -0.53 | (-0.64, -0.46) | -0.59 (0.007) | -0.59 | $(-0.60,-0.59)$ |
| 4 | -0.56 (0.075) | -0.52 | $(-0.65,-0.46)$ | -0.57 (0.078) | -0.56 | $(-0.67,-0.47)$ |
| 5 | -0.54 (0.039) | -0.54 | (-0.59, -0.49) | -0.55 (0.025) | -0.56 | $(-0.58,-0.52)$ |
| 6 | -0.56 (0.031) | -0.56 | $(-0.60,-0.52)$ | -0.55 (0.052) | -0.56 | $(-0.61,-0.48)$ |
| 7 | -0.55 (0.070) | -0.55 | $(-0.63,-0.46)$ | -0.57 (0.018) | -0.56 | $(-0.59,-0.55)$ |
| 8 | -0.63 (0.044) | -0.62 | $(-0.68,-0.57)$ | -0.59 (0.060) | -0.58 | $(-0.66,-0.51)$ |
| 9 | -0.60 (0.099) | -0.60 | $(-0.72,-0.48)$ | -0.59 (0.029) | -0.58 | $(-0.62,-0.55)$ |
| 10 | -0.56 (0.031) | -0.54 | (-0.60, -0.52) | -0.58 (0.057) | -0.58 | $(-0.66,-0.51)$ |
| 11 | -0.55 (0.047) | -0.56 | (-0.61, -0.49) | -0.60 (0.084) | -0.59 | ( $-0.70,-0.49$ ) |
| 12 | -0.59 (0.059) | -0.60 | (-0.67, -0.52) | -0.60 (0.068) | -0.61 | $(-0.68,-0.51)$ |
| 13 | -0.56 (0.047) | -0.56 | $(-0.62,-0.50)$ | -0.44 (0.412) | -0.59 | $(-0.95,0.07)$ |
| 14 | -0.54 (0.039) | -0.57 | $(-0.59,-0.50)$ | -0.40 (0.384) | -0.56 | $(-0.88,0.08)$ |
| 15 | -0.51 (0.071) | -0.48 | (-0.60, -0.42) | -0.60 (0.057) | -0.60 | $(-0.67,-0.53)$ |
| 16 | -0.59 (0.043) | -0.57 | $(-0.65,-0.54)$ | -0.55 (0.020) | -0.54 | $(-0.57,-0.52)$ |
| 17 | -0.60 (0.033) | -0.62 | (-0.64, -0.56) | -0.62 (0.050) | -0.61 | $(-0.68,-0.55)$ |
| 18 | -0.58 (0.060) | -0.57 | (-0.66, -0.51) | -0.60 (0.076) | -0.63 | $(-0.70,-0.51)$ |
| 19 | -0.57 (0.074) | -0.60 | $(-0.66,-0.48)$ | -0.57 (0.057) | -0.58 | $(-0.64,-0.50)$ |
| 20 | -0.60 (0.049) | -0.60 | $(-0.67,-0.54)$ | $-0.54(0.076)$ | -0.51 | $(-0.64,-0.45)$ |
| 21 | -0.61 (0.077) | -0.62 | $(-0.70,-0.51)$ | -0.60 (0.045) | -0.60 | $(-0.65,-0.54)$ |
| 22 | -0.54 (0.063) | -0.53 | (-0.62, -0.46) | -0.58 (0.044) | -0.60 | (-0.63, -0.52) |
| 23 | -0.55 (0.028) | -0.55 | $(-0.59,-0.52)$ | -0.56 (0.061) | -0.56 | $(-0.64,-0.49)$ |
| 24 | -0.57 (0.032) | -0.58 | (-0.61, -0.53) | -0.58 (0.088) | -0.55 | $(-0.69,-0.47)$ |
| 25 | -0.57 (0.072) | -0.56 | $(-0.66,-0.48)$ | -0.57 (0.057) | -0.60 | $(-0.64,-0.50)$ |
| 26 | -0.57 (0.037) | -0.59 | $(-0.61,-0.52)$ | -0.55 (0.029) | -0.54 | $(-0.58,-0.51)$ |
| 27 | -0.57 (0.079) | -0.59 | $(-0.66,-0.47)$ | -0.54 (0.042) | -0.53 | $(-0.59,-0.49)$ |
| 28 | -0.55 (0.066) | -0.56 | (-0.64, -0.47) | -0.36 (0.369) | -0.52 | $(-0.82,0.10)$ |
| 29 | -0.59 (0.032) | -0.59 | $(-0.63,-0.55)$ | -0.55 (0.034) | -0.55 | $(-0.59,-0.51)$ |
| 30 | -0.38 (0.358) | -0.53 | (-0.82, 0.07) | -0.58 (0.059) | -0.55 | $(-0.65,-0.51)$ |

Table J-2. Final Discrimination Parameter for $n=30$ Patients

| $\mathbf{y}$ AE | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | 95\% CI <br> on Mean | Mean (SD) | Median | 95\% CI <br> on Mean |
|  | $0.99(0.245)$ | 0.99 | $(0.97,1.02)$ | $0.99(0.249)$ | 0.98 | $(0.96,1.01)$ |
| 2 | $0.98(0.239)$ | 0.97 | $(0.95,1.00)$ | $0.99(0.249)$ | 0.99 | $(0.96,1.01)$ |
| 3 | $0.98(0.243)$ | 0.98 | $(0.96,1.01)$ | $0.99(0.242)$ | 0.99 | $(0.96,1.01)$ |
| 4 | $0.97(0.238)$ | 0.97 | $(0.95,0.99)$ | $0.99(0.242)$ | 0.98 | $(0.96,1.01)$ |
| 5 | $0.98(0.239)$ | 0.98 | $(0.96,1.01)$ | $0.98(0.249)$ | 0.98 | $(0.96,1.01)$ |
| 6 | $0.98(0.247)$ | 0.98 | $(0.96,1.01)$ | $0.99(0.251)$ | 0.99 | $(0.96,1.01)$ |
| 7 | $0.99(0.245)$ | 0.98 | $(0.96,1.01)$ | $0.98(0.247)$ | 0.98 | $(0.95,1.00)$ |
| 8 | $0.98(0.252)$ | 0.98 | $(0.95,1.00)$ | $0.97(0.243)$ | 0.97 | $(0.95,1.00)$ |
| 9 | $0.99(0.250)$ | 0.99 | $(0.96,1.02)$ | $0.99(0.246)$ | 0.99 | $(0.96,1.01)$ |

Table J-3. Final Difficulty Parameter for $n=30$ Patients

| $\mathbf{y} \mathbf{A} \mathbf{A E}$ | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | 95\% CI <br> on Mean | Mean (SD) | Median | 95\% CI <br> on Mean |
|  | $1.82(0.495)$ | 1.78 | $(1.75,1.90)$ | $1.76(0.501)$ | 1.74 | $(1.69,1.84)$ |
| 2 | $1.90(0.532)$ | 1.88 | $(1.82,1.98)$ | $1.97(0.547)$ | 1.92 | $(1.88,2.05)$ |
| 3 | $1.92(0.536)$ | 1.89 | $(1.83,2.00)$ | $1.83(0.519)$ | 1.80 | $(1.75,1.91)$ |
| 4 | $1.89(0.548)$ | 1.85 | $(1.80,1.97)$ | $1.91(0.545)$ | 1.86 | $(1.82,1.99)$ |
| 5 | $1.94(0.513)$ | 1.90 | $(1.86,2.01)$ | $1.88(0.529)$ | 1.86 | $(1.80,1.96)$ |
| 6 | $1.89(0.538)$ | 1.86 | $(1.81,1.98)$ | $1.87(0.548)$ | 1.83 | $(1.79,1.96)$ |
| 7 | $1.91(0.538)$ | 1.88 | $(1.83,1.99)$ | $1.92(0.558)$ | 1.88 | $(1.83,2.01)$ |
| 8 | $1.94(0.554)$ | 1.86 | $(1.85,2.02)$ | $1.93(0.542)$ | 1.89 | $(1.84,2.01)$ |
| 9 | $1.92(0.533)$ | 1.90 | $(1.84,2.00)$ | $1.92(0.572)$ | 1.88 | $(1.83,2.01)$ |

Table J-4. Final AE Predisposition Parameter for $n=250$ Patients

| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{gathered} \text { 95\% CI } \\ \text { on Mean } \end{gathered}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 1 | -0.05 (0.240) | -0.08 | ( $-0.12,0.02$ ) | -0.04 (0.069) | -0.05 | $(-0.12,0.05)$ |
| 2 | -0.04 (0.297) | -0.08 | ( $-0.12,0.04$ ) | -0.11 (0.147) | -0.06 | $(-0.30,0.07)$ |
| 3 | -0.11 (0.114) | -0.11 | $(-0.15,-0.08)$ | -0.21 (0.157) | -0.17 | $(-0.40,-0.02)$ |
| 4 | -0.11 (0.131) | -0.08 | $(-0.15,-0.08)$ | -0.16 (0.173) | -0.13 | $(-0.38,0.05)$ |
| 5 | -0.09 (0.130) | -0.08 | $(-0.12,-0.05)$ | -0.21 (0.116) | -0.18 | $(-0.35,-0.06)$ |
| 6 | -0.06 (0.218) | -0.06 | ( $-0.12,0.00$ ) | -0.07 (0.143) | -0.02 | $(-0.25,0.10)$ |
| 7 | -0.12 (0.133) | -0.13 | $(-0.16,-0.08)$ | -0.27 (0.095) | -0.27 | $(-0.39,-0.15)$ |
| 8 | -0.06 (0.226) | -0.07 | $(-0.13,0.00)$ | -0.13 (0.130) | -0.17 | $(-0.29,0.03)$ |
| 9 | -0.09 (0.212) | -0.09 | $(-0.15,-0.03)$ | -0.09 (0.102) | -0.11 | $(-0.22,0.04)$ |
| 10 | -0.14 (0.179) | -0.10 | $(-0.19,-0.09)$ | 0.08 (0.698) | -0.21 | $(-0.78,0.95)$ |
| 11 | -0.08 (0.247) | -0.09 | $(-0.15,-0.01)$ | -0.17 (0.118) | -0.20 | (-0.32, -0.03) |
| 12 | -0.11 (0.113) | -0.12 | $(-0.14,-0.08)$ | -0.17 (0.176) | -0.22 | $(-0.38,0.05)$ |
| 13 | -0.09 (0.261) | -0.12 | $(-0.16,-0.01)$ | -0.16 (0.243) | -0.23 | $(-0.46,0.15)$ |
| 14 | -0.03 (0.326) | -0.09 | ( $-0.12,0.07$ ) | -0.07 (0.102) | -0.03 | $(-0.19,0.06)$ |
| 15 | -0.09 (0.125) | -0.11 | $(-0.13,-0.06)$ | -0.08 (0.094) | -0.12 | $(-0.20,0.03)$ |
| 16 | -0.07 (0.241) | -0.09 | ( $-0.13,0.00$ ) | -0.11 (0.137) | -0.09 | $(-0.28,0.06)$ |
| 17 | -0.08 (0.122) | -0.08 | $(-0.12,-0.05)$ | -0.17 (0.094) | -0.19 | $(-0.29,-0.05)$ |
| 18 | -0.09 (0.150) | -0.09 | $(-0.13,-0.05)$ | -0.13 (0.094) | -0.13 | (-0.24, -0.01) |
| 19 | -0.07 (0.223) | -0.10 | $(-0.14,-0.01)$ | -0.08 (0.102) | -0.06 | $(-0.21,0.05)$ |
| 20 | -0.07 (0.241) | -0.10 | $(-0.14,-0.00)$ | -0.09 (0.078) | -0.08 | $(-0.19,0.00)$ |
| 21 | -0.11 (0.121) | -0.11 | $(-0.15,-0.08)$ | -0.06 (0.044) | -0.06 | $(-0.11,-0.00)$ |
| 22 | -0.12 (0.118) | -0.12 | $(-0.16,-0.09)$ | -0.03 (0.117) | -0.00 | $(-0.18,0.12)$ |
| 23 | -0.10 (0.261) | -0.13 | $(-0.17,-0.03)$ | -0.04 (0.213) | -0.04 | $(-0.31,0.22)$ |
| 24 | -0.08(0.245) | -0.10 | $(-0.15,-0.01)$ | -0.18 (0.123) | -0.14 | $(-0.34,-0.03)$ |
| 25 | -0.05 (0.280) | -0.09 | $(-0.13,0.02)$ | -0.16 (0.134) | -0.13 | $(-0.33,0.01)$ |
| 26 | -0.09 (0.134) | -0.09 | $(-0.13,-0.05)$ | -0.14 (0.099) | -0.13 | $(-0.26,-0.01)$ |
| 27 | -0.10 (0.143) | -0.09 | $(-0.14,-0.05)$ | -0.09 (0.046) | -0.11 | $(-0.14,-0.03)$ |
| 28 | -0.12 (0.124) | -0.13 | $(-0.15,-0.08)$ | -0.23 (0.140) | -0.24 | $(-0.40,-0.05)$ |
| 29 | -0.05 (0.311) | -0.09 | $(-0.14,0.04)$ | -0.05 (0.135) | -0.11 | $(-0.22,0.11)$ |
| 30 | -0.10 (0.233) | -0.14 | $(-0.17,-0.04)$ | -0.11 (0.119) | -0.14 | $(-0.26,0.04)$ |
| 31 | -0.11 (0.212) | -0.14 | $(-0.17,-0.05)$ | -0.09 (0.079) | -0.07 | $(-0.19,0.01)$ |
| 32 | -0.09 (0.108) | -0.11 | $(-0.12,-0.06)$ | 0.03 (0.137) | 0.07 | $(-0.14,0.20)$ |
| 33 | -0.11 (0.153) | -0.10 | $(-0.15,-0.06)$ | -0.13 (0.140) | -0.11 | $(-0.30,0.05)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 34 | -0.10 (0.134) | -0.12 | (-0.14, -0.07) | -0.07 (0.177) | -0.13 | $(-0.29,0.15)$ |
| 35 | -0.10 (0.131) | -0.08 | (-0.14, -0.06) | -0.06 (0.129) | -0.10 | $(-0.22,0.10)$ |
| 36 | -0.05 (0.348) | -0.12 | $(-0.15,0.05)$ | -0.14 (0.166) | -0.20 | ( $-0.34,0.07$ ) |
| 37 | -0.10 (0.147) | -0.13 | (-0.14, -0.06) | -0.11 (0.035) | -0.11 | $(-0.15,-0.06)$ |
| 38 | -0.09 (0.123) | -0.12 | (-0.13, -0.06) | -0.17 (0.150) | -0.21 | $(-0.36,0.02)$ |
| 39 | -0.05 (0.296) | -0.11 | $(-0.13,0.04)$ | -0.08 (0.094) | -0.11 | $(-0.19,0.04)$ |
| 40 | -0.14 (0.146) | -0.13 | $(-0.18,-0.09)$ | -0.01 (0.115) | 0.01 | $(-0.15,0.14)$ |
| 41 | -0.10 (0.228) | -0.13 | $(-0.17,-0.04)$ | -0.02 (0.214) | -0.02 | $(-0.29,0.24)$ |
| 42 | -0.10 (0.145) | -0.11 | $(-0.15,-0.06)$ | -0.15 (0.099) | -0.18 | (-0.27, -0.03) |
| 43 | -0.08 (0.132) | -0.08 | (-0.12, -0.05) | -0.13 (0.206) | -0.13 | $(-0.38,0.13)$ |
| 44 | -0.08 (0.231) | -0.10 | $(-0.14,-0.01)$ | -0.19 (0.128) | -0.23 | $(-0.35,-0.03)$ |
| 45 | -0.10 (0.149) | -0.11 | $(-0.14,-0.06)$ | -0.02 (0.138) | -0.06 | $(-0.19,0.15)$ |
| 46 | -0.09 (0.231) | -0.10 | $(-0.16,-0.02)$ | -0.13 (0.179) | -0.12 | $(-0.35,0.10)$ |
| 47 | -0.11 (0.134) | -0.12 | $(-0.14,-0.07)$ | $-0.11(0.051)$ | -0.11 | (-0.17, -0.04) |
| 48 | -0.06 (0.248) | -0.07 | $(-0.13,0.02)$ | -0.10 (0.112) | -0.10 | $(-0.24,0.04)$ |
| 49 | -0.12 (0.117) | -0.09 | $(-0.15,-0.09)$ | 0.15 (0.661) | -0.08 | $(-0.67,0.97)$ |
| 50 | -0.11 (0.160) | -0.09 | $(-0.16,-0.07)$ | -0.16 (0.246) | -0.25 | $(-0.47,0.14)$ |
| 51 | -0.04 (0.308) | -0.10 | $(-0.13,0.05)$ | 0.06 (0.521) | -0.15 | $(-0.58,0.71)$ |
| 52 | -0.07 (0.125) | -0.05 | $(-0.11,-0.04)$ | -0.13 (0.102) | -0.19 | $(-0.26,-0.00)$ |
| 53 | -0.12 (0.151) | -0.11 | (-0.16, -0.08) | -0.01 (0.045) | -0.01 | $(-0.07,0.05)$ |
| 54 | -0.11 (0.215) | -0.14 | (-0.18, -0.05) | -0.18 (0.093) | -0.16 | $(-0.29,-0.06)$ |
| 55 | -0.06 (0.255) | -0.11 | $(-0.13,0.01)$ | -0.05 (0.075) | -0.06 | $(-0.15,0.04)$ |
| 56 | -0.10 (0.231) | -0.13 | $(-0.17,-0.04)$ | -0.07 (0.066) | -0.04 | $(-0.15,0.02)$ |
| 57 | -0.07 (0.203) | -0.09 | $(-0.13,-0.01)$ | -0.07 (0.113) | -0.08 | $(-0.21,0.07)$ |
| 58 | -0.11 (0.144) | -0.12 | $(-0.15,-0.07)$ | -0.14 (0.132) | -0.09 | $(-0.30,0.03)$ |
| 59 | -0.11 (0.158) | -0.10 | $(-0.16,-0.07)$ | -0.13 (0.154) | -0.16 | $(-0.32,0.06)$ |
| 60 | -0.11 (0.248) | -0.12 | $(-0.18,-0.04)$ | -0.15 (0.116) | -0.17 | $(-0.29,-0.00)$ |
| 61 | -0.05 (0.233) | -0.05 | $(-0.12,0.01)$ | -0.14 (0.092) | -0.14 | $(-0.25,-0.02)$ |
| 62 | -0.11 (0.118) | -0.10 | (-0.15, -0.08) | -0.01 (0.243) | -0.06 | $(-0.31,0.29)$ |
| 63 | -0.06 (0.279) | -0.11 | (-0.14, 0.02) | -0.13 (0.121) | -0.09 | $(-0.28,0.02)$ |
| 64 | -0.06 (0.234) | -0.08 | $(-0.13,0.00)$ | -0.17 (0.115) | -0.18 | $(-0.32,-0.03)$ |
| 65 | -0.10 (0.125) | -0.10 | $(-0.14,-0.07)$ | -0.14 (0.118) | -0.14 | $(-0.29,0.01)$ |
| 66 | -0.10 (0.228) | -0.13 | (-0.17, -0.04) | -0.06 (0.223) | -0.12 | $(-0.33,0.22)$ |
| 67 | -0.10 (0.143) | -0.10 | (-0.14, -0.06) | -0.12 (0.180) | -0.05 | $(-0.34,0.11)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 68 | -0.04 (0.220) | -0.08 | (-0.10, 0.02) | -0.05 (0.081) | -0.03 | ( $-0.15,0.05$ ) |
| 69 | -0.14 (0.124) | -0.13 | $(-0.18,-0.11)$ | -0.18 (0.129) | -0.17 | ( $-0.34,-0.02$ ) |
| 70 | -0.13 (0.139) | -0.16 | ( $-0.17,-0.09$ ) | -0.21 (0.126) | -0.26 | $(-0.37,-0.05)$ |
| 71 | -0.10 (0.133) | -0.09 | $(-0.13,-0.06)$ | -0.06 (0.107) | -0.08 | $(-0.19,0.07)$ |
| 72 | -0.05 (0.283) | -0.08 | $(-0.13,0.03)$ | 0.01 (0.110) | -0.02 | $(-0.13,0.14)$ |
| 73 | -0.07 (0.237) | -0.09 | (-0.14, -0.00) | -0.10 (0.098) | -0.11 | ( $-0.22,0.02$ ) |
| 74 | -0.10 (0.124) | -0.10 | $(-0.13,-0.06)$ | -0.25 (0.120) | -0.32 | $(-0.40,-0.10)$ |
| 75 | -0.04 (0.362) | -0.11 | $(-0.14,0.07)$ | -0.01 (0.073) | 0.03 | $(-0.10,0.08)$ |
| 76 | -0.09 (0.128) | -0.11 | (-0.12, -0.05) | -0.13 (0.215) | -0.14 | ( $-0.40,0.14$ ) |
| 77 | -0.02 (0.418) | -0.13 | $(-0.13,0.10)$ | -0.03 (0.103) | -0.03 | $(-0.16,0.10)$ |
| 78 | -0.01 (0.378) | -0.09 | $(-0.12,0.10)$ | -0.09 (0.189) | -0.17 | $(-0.33,0.14)$ |
| 79 | -0.10 (0.135) | -0.12 | (-0.14, -0.06) | 0.12 (0.466) | -0.13 | ( $-0.46,0.70$ ) |
| 80 | -0.11 (0.133) | -0.13 | $(-0.15,-0.07)$ | -0.10 (0.140) | -0.09 | $(-0.28,0.07)$ |
| 81 | -0.12 (0.136) | -0.10 | $(-0.16,-0.08)$ | -0.18 (0.113) | -0.16 | $(-0.32,-0.04)$ |
| 82 | -0.08 (0.223) | -0.09 | $(-0.15,-0.02)$ | -0.10 (0.106) | -0.12 | $(-0.23,0.03)$ |
| 83 | -0.09 (0.128) | -0.10 | $(-0.13,-0.06)$ | -0.09 (0.126) | -0.12 | $(-0.25,0.06)$ |
| 84 | -0.04 (0.283) | -0.09 | $(-0.12,0.04)$ | -0.06 (0.083) | -0.08 | (-0.16, 0.04) |
| 85 | -0.09 (0.252) | -0.12 | $(-0.16,-0.02)$ | -0.11 (0.087) | -0.10 | (-0.22, 0.00) |
| 86 | -0.05 (0.291) | -0.08 | $(-0.14,0.03)$ | 0.11 (0.635) | -0.17 | $(-0.68,0.90)$ |
| 87 | -0.12 (0.144) | -0.08 | (-0.16, -0.08) | -0.09 (0.126) | -0.14 | $(-0.25,0.06)$ |
| 88 | -0.10 (0.236) | -0.14 | ( $-0.17,-0.03$ ) | -0.03 (0.182) | -0.02 | $(-0.25,0.20)$ |
| 89 | -0.06 (0.211) | -0.06 | (-0.12, -0.00) | -0.09 (0.131) | -0.11 | $(-0.26,0.07)$ |
| 90 | -0.10 (0.199) | -0.12 | $(-0.15,-0.04)$ | -0.10 (0.140) | -0.14 | (-0.27, 0.07) |
| 91 | -0.11 (0.123) | -0.13 | $(-0.15,-0.08)$ | 0.06 (0.519) | -0.13 | $(-0.58,0.71)$ |
| 92 | -0.01 (0.295) | -0.05 | $(-0.10,0.07)$ | 0.00 (0.188) | 0.04 | ( $-0.23,0.24$ ) |
| 93 | -0.08 (0.133) | -0.08 | $(-0.12,-0.04)$ | -0.16 (0.111) | -0.18 | $(-0.29,-0.02)$ |
| 94 | -0.08 (0.258) | -0.09 | $(-0.16,-0.01)$ | $-0.21(0.073)$ | -0.19 | $(-0.30,-0.12)$ |
| 95 | -0.07 (0.129) | -0.05 | $(-0.11,-0.04)$ | -0.04 (0.187) | -0.08 | $(-0.27,0.19)$ |
| 96 | -0.09 (0.234) | -0.12 | $(-0.15,-0.02)$ | -0.05 (0.116) | -0.01 | ( $-0.19,0.10$ ) |
| 97 | -0.16 (0.133) | -0.17 | $(-0.20,-0.12)$ | -0.12 (0.204) | -0.16 | $(-0.37,0.14)$ |
| 98 | -0.15 (0.136) | -0.16 | $(-0.19,-0.11)$ | 0.02 (0.120) | -0.02 | $(-0.13,0.16)$ |
| 99 | -0.08 (0.120) | -0.08 | $(-0.11,-0.04)$ | -0.17 (0.210) | -0.06 | $(-0.43,0.10)$ |
| 100 | -0.10 (0.121) | -0.08 | $(-0.13,-0.06)$ | -0.11 (0.186) | -0.07 | ( $-0.34,0.12$ ) |
| 101 | -0.08 (0.229) | -0.11 | $(-0.15,-0.02)$ | -0.14 (0.133) | -0.07 | $(-0.30,0.03)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 102 | -0.09 (0.131) | -0.11 | $(-0.13,-0.05)$ | -0.06 (0.061) | -0.08 | ( $-0.14,0.02$ ) |
| 103 | -0.06 (0.128) | -0.05 | $(-0.10,-0.02)$ | -0.08 (0.066) | -0.07 | $(-0.16,0.00)$ |
| 104 | -0.04 (0.319) | -0.08 | ( $-0.13,0.05$ ) | -0.19 (0.101) | -0.19 | $(-0.31,-0.06)$ |
| 105 | -0.13 (0.147) | -0.12 | $(-0.18,-0.09)$ | -0.09 (0.157) | -0.09 | (-0.29, 0.10) |
| 106 | -0.12 (0.130) | -0.11 | $(-0.16,-0.09)$ | -0.09 (0.032) | -0.10 | $(-0.13,-0.05)$ |
| 107 | -0.05 (0.299) | -0.12 | $(-0.13,0.04)$ | -0.06 (0.133) | -0.06 | $(-0.23,0.11)$ |
| 108 | -0.10 (0.250) | -0.14 | $(-0.17,-0.03)$ | -0.13 (0.108) | -0.18 | $(-0.26,0.01)$ |
| 109 | -0.15 (0.124) | -0.17 | $(-0.18,-0.11)$ | -0.05 (0.100) | -0.05 | ( $-0.17,0.08$ ) |
| 110 | -0.07 (0.137) | -0.09 | $(-0.11,-0.03)$ | -0.15 (0.193) | -0.22 | $(-0.39,0.09)$ |
| 111 | -0.12 (0.127) | -0.15 | $(-0.16,-0.09)$ | -0.13 (0.061) | -0.11 | $(-0.20,-0.05)$ |
| 112 | -0.08 (0.253) | -0.13 | $(-0.15,-0.01)$ | -0.17 (0.079) | -0.16 | $(-0.27,-0.07)$ |
| 113 | -0.10 (0.107) | -0.09 | $(-0.14,-0.07)$ | -0.06 (0.123) | -0.09 | ( $-0.21,0.09$ ) |
| 114 | -0.11 (0.105) | -0.10 | $(-0.14,-0.08)$ | -0.23 (0.079) | -0.24 | $(-0.32,-0.13)$ |
| 115 | -0.09 (0.121) | -0.09 | $(-0.12,-0.05)$ | -0.15 (0.065) | -0.15 | $(-0.23,-0.07)$ |
| 116 | -0.09 (0.124) | -0.08 | $(-0.12,-0.05)$ | -0.02 (0.171) | -0.07 | $(-0.24,0.19)$ |
| 117 | -0.13 (0.123) | -0.15 | $(-0.17,-0.10)$ | -0.11 (0.211) | -0.17 | ( $-0.37,0.15$ ) |
| 118 | -0.07 (0.227) | -0.10 | $(-0.13,-0.00)$ | -0.04 (0.120) | -0.07 | $(-0.19,0.11)$ |
| 119 | -0.11 (0.138) | -0.11 | $(-0.15,-0.07)$ | -0.10 (0.176) | -0.07 | $(-0.32,0.12)$ |
| 120 | -0.07 (0.144) | -0.07 | $(-0.11,-0.03)$ | -0.09 (0.148) | -0.08 | $(-0.28,0.09)$ |
| 121 | -0.08 (0.216) | -0.09 | $(-0.14,-0.01)$ | -0.13 (0.123) | -0.09 | $(-0.28,0.02)$ |
| 122 | -0.16 (0.102) | -0.18 | $(-0.18,-0.13)$ | -0.09 (0.214) | -0.15 | $(-0.35,0.18)$ |
| 123 | -0.10 (0.139) | -0.10 | $(-0.14,-0.06)$ | -0.14 (0.156) | -0.13 | $(-0.33,0.05)$ |
| 124 | -0.08 (0.145) | -0.06 | $(-0.12,-0.03)$ | -0.09 (0.097) | -0.06 | (-0.21, 0.03) |
| 125 | -0.11 (0.135) | -0.12 | $(-0.15,-0.07)$ | -0.07 (0.094) | -0.09 | $(-0.19,0.04)$ |
| 126 | -0.06 (0.235) | -0.10 | $(-0.12,0.01)$ | -0.15 (0.104) | -0.12 | $(-0.28,-0.02)$ |
| 127 | -0.14 (0.111) | -0.16 | $(-0.17,-0.11)$ | -0.04 (0.076) | -0.02 | $(-0.13,0.06)$ |
| 128 | -0.09 (0.204) | -0.10 | $(-0.15,-0.03)$ | -0.17 (0.211) | -0.12 | $(-0.43,0.09)$ |
| 129 | -0.08 (0.319) | -0.12 | $(-0.17,0.01)$ | -0.03 (0.122) | -0.03 | $(-0.18,0.12)$ |
| 130 | -0.14 (0.148) | -0.13 | $(-0.18,-0.10)$ | -0.11 (0.165) | -0.12 | $(-0.32,0.09)$ |
| 131 | -0.11 (0.140) | -0.13 | $(-0.15,-0.07)$ | -0.15 (0.118) | -0.19 | $(-0.30,-0.00)$ |
| 132 | -0.10 (0.307) | -0.15 | $(-0.19,-0.01)$ | -0.22 (0.145) | -0.29 | $(-0.40,-0.04)$ |
| 133 | -0.08 (0.213) | -0.11 | $(-0.14,-0.02)$ | -0.06 (0.046) | -0.04 | $(-0.12,-0.01)$ |
| 134 | -0.01 (0.352) | -0.10 | ( $-0.11,0.09$ ) | -0.13 (0.112) | -0.09 | ( $-0.27,0.01$ ) |
| 135 | -0.10 (0.131) | -0.10 | $(-0.14,-0.06)$ | -0.13 (0.164) | -0.17 | $(-0.33,0.08)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 136 | -0.10 (0.129) | -0.10 | $(-0.13,-0.06)$ | -0.05 (0.096) | -0.03 | $(-0.17,0.07)$ |
| 137 | -0.03 (0.325) | -0.07 | ( $-0.12,0.06$ ) | -0.10 (0.067) | -0.09 | $(-0.18,-0.01)$ |
| 138 | -0.08 (0.128) | -0.09 | $(-0.11,-0.04)$ | -0.03 (0.090) | -0.04 | $(-0.15,0.08)$ |
| 139 | -0.09 (0.186) | -0.12 | $(-0.14,-0.03)$ | -0.06 (0.154) | -0.04 | $(-0.25,0.13)$ |
| 140 | -0.07 (0.219) | -0.07 | $(-0.13,-0.01)$ | -0.20 (0.040) | -0.17 | $(-0.24,-0.15)$ |
| 141 | -0.10 (0.133) | -0.09 | $(-0.14,-0.06)$ | -0.06 (0.093) | -0.06 | $(-0.18,0.05)$ |
| 142 | -0.10 (0.299) | -0.14 | $(-0.19,-0.02)$ | -0.06 (0.127) | -0.00 | $(-0.22,0.10)$ |
| 143 | -0.12 (0.123) | -0.13 | $(-0.16,-0.09)$ | -0.10 (0.113) | -0.11 | ( $-0.24,0.04$ ) |
| 144 | -0.07 (0.142) | -0.07 | $(-0.11,-0.03)$ | -0.14 (0.111) | -0.13 | $(-0.28,-0.01)$ |
| 145 | -0.09 (0.220) | -0.11 | $(-0.15,-0.02)$ | 0.15 (0.636) | -0.11 | $(-0.64,0.94)$ |
| 146 | -0.07 (0.268) | -0.11 | $(-0.15,0.01)$ | -0.05 (0.206) | -0.03 | $(-0.30,0.21)$ |
| 147 | -0.12 (0.122) | -0.12 | $(-0.16,-0.09)$ | -0.06 (0.173) | -0.03 | $(-0.28,0.15)$ |
| 148 | -0.09 (0.231) | -0.13 | $(-0.15,-0.02)$ | -0.02 (0.098) | 0.01 | ( $-0.14,0.10$ ) |
| 149 | -0.10 (0.115) | -0.10 | $(-0.14,-0.07)$ | -0.07 (0.045) | -0.08 | $(-0.12,-0.01)$ |
| 150 | -0.13 (0.123) | -0.13 | $(-0.16,-0.09)$ | -0.15 (0.157) | -0.08 | $(-0.35,0.04)$ |
| 151 | -0.05 (0.287) | -0.09 | ( $-0.13,0.03$ ) | -0.19 (0.161) | -0.11 | $(-0.39,0.01)$ |
| 152 | -0.05 (0.222) | -0.07 | $(-0.11,0.01)$ | -0.07 (0.098) | -0.10 | (-0.19, 0.05) |
| 153 | -0.11 (0.123) | -0.11 | $(-0.15,-0.08)$ | 0.11 (0.600) | -0.08 | ( $-0.64,0.85$ ) |
| 154 | -0.03 (0.320) | -0.11 | $(-0.12,0.06)$ | -0.20 (0.039) | -0.20 | $(-0.25,-0.15)$ |
| 155 | -0.12 (0.127) | -0.12 | $(-0.16,-0.09)$ | -0.02 (0.062) | -0.06 | $(-0.10,0.06)$ |
| 156 | -0.07 (0.130) | -0.07 | $(-0.11,-0.04)$ | -0.14 (0.071) | -0.17 | $(-0.23,-0.05)$ |
| 157 | -0.08 (0.130) | -0.06 | $(-0.12,-0.04)$ | 0.16 (0.554) | -0.05 | $(-0.53,0.85)$ |
| 158 | -0.03 (0.289) | -0.08 | $(-0.12,0.05)$ | -0.13 (0.123) | -0.08 | $(-0.29,0.02)$ |
| 159 | -0.10 (0.233) | -0.13 | $(-0.17,-0.04)$ | -0.03 (0.125) | 0.05 | $(-0.18,0.13)$ |
| 160 | -0.10 (0.137) | -0.10 | $(-0.14,-0.06)$ | -0.10 (0.117) | -0.07 | $(-0.25,0.04)$ |
| 161 | -0.10 (0.126) | -0.12 | $(-0.13,-0.06)$ | -0.23 (0.091) | -0.22 | $(-0.34,-0.12)$ |
| 162 | -0.11 (0.119) | -0.10 | (-0.14, -0.07) | -0.10 (0.123) | -0.16 | $(-0.25,0.05)$ |
| 163 | -0.04 (0.248) | -0.07 | $(-0.11,0.03)$ | -0.15 (0.111) | -0.12 | $(-0.29,-0.01)$ |
| 164 | -0.15 (0.145) | -0.14 | $(-0.19,-0.11)$ | -0.06 (0.123) | -0.04 | (-0.22, 0.09) |
| 165 | -0.09 (0.125) | -0.08 | $(-0.12,-0.05)$ | -0.09 (0.129) | -0.15 | $(-0.25,0.07)$ |
| 166 | -0.11 (0.140) | -0.11 | $(-0.15,-0.07)$ | -0.14 (0.123) | -0.16 | $(-0.29,0.01)$ |
| 167 | -0.13 (0.120) | -0.14 | $(-0.17,-0.10)$ | 0.21 (0.621) | -0.10 | $(-0.56,0.98)$ |
| 168 | -0.09 (0.222) | -0.12 | $(-0.16,-0.03)$ | -0.01 (0.207) | 0.09 | ( $-0.27,0.25$ ) |
| 169 | -0.11 (0.124) | -0.09 | $(-0.15,-0.08)$ | -0.09 (0.073) | -0.11 | $(-0.18,0.00)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 170 | -0.10 (0.130) | -0.11 | $(-0.14,-0.07)$ | -0.22 (0.056) | -0.19 | $(-0.28,-0.15)$ |
| 171 | -0.06 (0.147) | -0.09 | $(-0.10,-0.02)$ | -0.11 (0.082) | -0.13 | $(-0.21,-0.01)$ |
| 172 | -0.09 (0.120) | -0.11 | (-0.12, -0.05) | -0.13 (0.142) | -0.07 | ( $-0.30,0.05$ ) |
| 173 | -0.08 (0.307) | -0.12 | $(-0.16,0.01)$ | -0.12 (0.105) | -0.11 | $(-0.25,0.01)$ |
| 174 | -0.08 (0.231) | -0.10 | $(-0.15,-0.02)$ | 0.23 (0.584) | -0.01 | $(-0.49,0.96)$ |
| 175 | -0.08 (0.220) | -0.12 | $(-0.15,-0.02)$ | 0.16 (0.549) | -0.03 | ( $-0.52,0.84$ ) |
| 176 | -0.12 (0.236) | -0.14 | $(-0.19,-0.05)$ | -0.07 (0.095) | -0.04 | $(-0.19,0.05)$ |
| 177 | -0.08 (0.217) | -0.10 | (-0.14, -0.02) | -0.14 (0.068) | -0.12 | ( $-0.22,-0.06$ ) |
| 178 | -0.08 (0.244) | -0.10 | $(-0.15,-0.01)$ | 0.01 (0.222) | 0.13 | $(-0.27,0.29)$ |
| 179 | -0.13 (0.126) | -0.11 | (-0.17, -0.10) | -0.16 (0.115) | -0.21 | $(-0.30,-0.01)$ |
| 180 | -0.08 (0.239) | -0.08 | $(-0.15,-0.01)$ | -0.16 (0.136) | -0.17 | ( $-0.32,0.01$ ) |
| 181 | -0.08 (0.241) | -0.10 | (-0.14, -0.01) | -0.10 (0.114) | -0.08 | $(-0.25,0.04)$ |
| 182 | -0.06 (0.234) | -0.10 | (-0.13, 0.00) | 0.15 (0.604) | -0.07 | ( $-0.60,0.90$ ) |
| 183 | -0.03 (0.281) | -0.07 | (-0.11, 0.05) | -0.14 (0.067) | -0.15 | (-0.23, -0.06) |
| 184 | -0.12 (0.104) | -0.11 | $(-0.15,-0.09)$ | -0.12 (0.193) | -0.17 | $(-0.36,0.12)$ |
| 185 | -0.09 (0.149) | -0.09 | (-0.14, -0.05) | -0.15 (0.178) | -0.13 | ( $-0.37,0.07$ ) |
| 186 | -0.11 (0.132) | -0.14 | $(-0.15,-0.08)$ | -0.07 (0.070) | -0.09 | $(-0.16,0.02)$ |
| 187 | -0.11 (0.133) | -0.13 | ( $-0.15,-0.08$ ) | -0.10 (0.163) | -0.12 | $(-0.30,0.10)$ |
| 188 | -0.12 (0.124) | -0.13 | $(-0.15,-0.08)$ | -0.13 (0.158) | -0.11 | ( $-0.33,0.07$ ) |
| 189 | -0.12 (0.155) | -0.12 | $(-0.16,-0.07)$ | -0.08 (0.120) | -0.14 | ( $-0.23,0.07$ ) |
| 190 | -0.08 (0.245) | -0.11 | $(-0.15,-0.01)$ | 0.23 (0.582) | 0.00 | ( $-0.49,0.95$ ) |
| 191 | -0.12 (0.125) | -0.13 | $(-0.15,-0.08)$ | -0.12 (0.084) | -0.12 | $(-0.22,-0.01)$ |
| 192 | -0.10 (0.144) | -0.12 | $(-0.14,-0.06)$ | -0.10 (0.211) | 0.01 | $(-0.36,0.17)$ |
| 193 | -0.07 (0.210) | -0.10 | $(-0.13,-0.01)$ | -0.04 (0.126) | -0.02 | $(-0.20,0.12)$ |
| 194 | -0.09 (0.132) | -0.07 | $(-0.13,-0.05)$ | -0.06 (0.072) | -0.04 | $(-0.15,0.03)$ |
| 195 | -0.08 (0.223) | -0.08 | $(-0.14,-0.01)$ | -0.10 (0.091) | -0.13 | ( $-0.21,0.02$ ) |
| 196 | -0.10 (0.123) | -0.12 | $(-0.13,-0.06)$ | -0.06 (0.112) | -0.09 | $(-0.20,0.08)$ |
| 197 | -0.11 (0.144) | -0.12 | $(-0.15,-0.07)$ | -0.09 (0.122) | -0.10 | $(-0.25,0.06)$ |
| 198 | -0.10 (0.119) | -0.10 | $(-0.13,-0.06)$ | -0.10 (0.187) | -0.09 | $(-0.33,0.13)$ |
| 199 | -0.09 (0.201) | -0.10 | $(-0.14,-0.03)$ | -0.25 (0.143) | -0.22 | $(-0.43,-0.07)$ |
| 200 | -0.09 (0.203) | -0.10 | $(-0.14,-0.03)$ | -0.13 (0.164) | -0.14 | $(-0.34,0.07)$ |
| 201 | -0.12 (0.113) | -0.14 | $(-0.15,-0.08)$ | -0.13 (0.202) | -0.02 | $(-0.38,0.12)$ |
| 202 | -0.04 (0.248) | -0.09 | $(-0.11,0.03)$ | -0.08 (0.119) | -0.09 | $(-0.23,0.06)$ |
| 203 | -0.10 (0.125) | -0.08 | (-0.13, -0.06) | -0.05 (0.116) | -0.05 | $(-0.20,0.09)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ | Mean (SD) | Median | $\begin{aligned} & \text { 95\% CI } \\ & \text { on Mean } \end{aligned}$ |
| 204 | -0.10 (0.220) | -0.10 | $(-0.16,-0.03)$ | -0.07 (0.137) | -0.08 | ( $-0.24,0.10$ ) |
| 205 | -0.11 (0.126) | -0.11 | ( $-0.14,-0.07$ ) | -0.09 (0.219) | 0.04 | $(-0.36,0.18)$ |
| 206 | -0.10 (0.138) | -0.07 | ( $-0.14,-0.06$ ) | -0.02 (0.116) | -0.02 | ( $-0.16,0.13$ ) |
| 207 | -0.10 (0.121) | -0.08 | $(-0.13,-0.06)$ | -0.09 (0.061) | -0.12 | $(-0.17,-0.02)$ |
| 208 | -0.10 (0.231) | -0.12 | $(-0.17,-0.04)$ | 0.28 (0.636) | 0.03 | (-0.51, 1.06) |
| 209 | -0.10 (0.128) | -0.07 | ( $-0.13,-0.06$ ) | -0.13 (0.165) | -0.20 | $(-0.33,0.08)$ |
| 210 | -0.16 (0.153) | -0.15 | $(-0.20,-0.11)$ | -0.10 (0.088) | -0.10 | $(-0.21,0.01)$ |
| 211 | -0.00 (0.320) | -0.05 | $(-0.09,0.09)$ | -0.17 (0.216) | -0.15 | $(-0.44,0.10)$ |
| 212 | -0.11 (0.150) | -0.11 | $(-0.15,-0.07)$ | -0.04 (0.096) | -0.04 | $(-0.16,0.08)$ |
| 213 | -0.09 (0.241) | -0.12 | $(-0.15,-0.02)$ | -0.09 (0.065) | -0.08 | $(-0.17,-0.01)$ |
| 214 | -0.11 (0.210) | -0.15 | $(-0.17,-0.05)$ | -0.13 (0.094) | -0.19 | $(-0.24,-0.01)$ |
| 215 | -0.12 (0.139) | -0.13 | $(-0.16,-0.08)$ | -0.13 (0.101) | -0.14 | $(-0.26,-0.01)$ |
| 216 | -0.08 (0.140) | -0.09 | $(-0.12,-0.04)$ | -0.11 (0.093) | -0.09 | ( $-0.23,0.00$ ) |
| 217 | -0.11 (0.142) | -0.11 | $(-0.15,-0.07)$ | -0.08 (0.146) | -0.10 | $(-0.26,0.11)$ |
| 218 | -0.07 (0.259) | -0.08 | $(-0.14,0.00)$ | -0.10 (0.073) | -0.13 | $(-0.19,-0.01)$ |
| 219 | -0.09 (0.226) | -0.10 | $(-0.16,-0.03)$ | -0.18 (0.094) | -0.22 | $(-0.30,-0.07)$ |
| 220 | -0.06 (0.282) | -0.12 | $(-0.14,0.02)$ | -0.06 (0.083) | -0.08 | $(-0.17,0.04)$ |
| 221 | -0.09 (0.137) | -0.11 | $(-0.13,-0.05)$ | -0.13 (0.140) | -0.19 | ( $-0.30,0.05$ ) |
| 222 | -0.12 (0.118) | -0.11 | $(-0.15,-0.08)$ | -0.07 (0.131) | -0.09 | $(-0.23,0.09)$ |
| 223 | -0.06 (0.219) | -0.07 | $(-0.12,-0.00)$ | 0.02 (0.101) | -0.00 | (-0.11, 0.14) |
| 224 | -0.11 (0.117) | -0.10 | $(-0.15,-0.08)$ | -0.13 (0.205) | -0.15 | $(-0.39,0.12)$ |
| 225 | -0.08 (0.236) | -0.11 | $(-0.15,-0.02)$ | -0.07 (0.142) | -0.05 | $(-0.25,0.11)$ |
| 226 | -0.07 (0.209) | -0.10 | $(-0.13,-0.01)$ | -0.07 (0.193) | -0.07 | (-0.31, 0.17) |
| 227 | -0.11 (0.128) | -0.09 | $(-0.15,-0.07)$ | -0.13 (0.185) | -0.06 | $(-0.36,0.10)$ |
| 228 | -0.11 (0.238) | -0.13 | $(-0.18,-0.05)$ | -0.07 (0.084) | -0.04 | $(-0.17,0.04)$ |
| 229 | -0.04 (0.244) | -0.07 | $(-0.11,0.03)$ | -0.03 (0.184) | -0.04 | $(-0.26,0.20)$ |
| 230 | -0.07 (0.242) | -0.11 | $(-0.14,-0.00)$ | -0.06 (0.184) | -0.08 | $(-0.29,0.17)$ |
| 231 | -0.05 (0.304) | -0.09 | $(-0.14,0.03)$ | -0.15 (0.052) | -0.15 | $(-0.21,-0.09)$ |
| 232 | -0.10 (0.256) | -0.12 | $(-0.17,-0.03)$ | -0.10 (0.110) | -0.08 | $(-0.24,0.03)$ |
| 233 | -0.11 (0.124) | -0.11 | $(-0.15,-0.08)$ | -0.13 (0.079) | -0.16 | $(-0.23,-0.03)$ |
| 234 | -0.09 (0.131) | -0.07 | ( $-0.13,-0.06$ ) | -0.15 (0.138) | -0.13 | ( $-0.32,0.02$ ) |
| 235 | -0.09 (0.216) | -0.11 | $(-0.16,-0.03)$ | 0.01 (0.114) | 0.05 | $(-0.14,0.15)$ |
| 236 | -0.13 (0.140) | -0.13 | ( $-0.17,-0.09$ ) | -0.11 (0.081) | -0.16 | $(-0.21,-0.01)$ |
| 237 | -0.11 (0.124) | -0.13 | $(-0.15,-0.08)$ | 0.25 (0.684) | -0.06 | $(-0.60,1.10)$ |


| Patient | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | 95\% CI <br> on Mean | Mean (SD) | Median | 95\% CI <br> on Mean |
|  | $-0.10(0.119)$ | -0.11 | $(-0.13,-0.06)$ | $-0.12(0.097)$ | -0.14 | $(-0.24,-0.00)$ |
| 239 | $-0.13(0.147)$ | -0.12 | $(-0.17,-0.09)$ | $-0.06(0.272)$ | -0.12 | $(-0.40,0.27)$ |
| 240 | $-0.08(0.237)$ | -0.11 | $(-0.15,-0.01)$ | $-0.10(0.069)$ | -0.12 | $(-0.19,-0.02)$ |
| 241 | $-0.09(0.208)$ | -0.09 | $(-0.15,-0.03)$ | $-0.07(0.073)$ | -0.04 | $(-0.16,0.02)$ |
| 242 | $-0.09(0.144)$ | -0.09 | $(-0.13,-0.05)$ | $-0.04(0.251)$ | 0.04 | $(-0.35,0.28)$ |
| 243 | $-0.11(0.146)$ | -0.11 | $(-0.15,-0.07)$ | $-0.10(0.101)$ | -0.12 | $(-0.22,0.03)$ |
| 244 | $-0.06(0.215)$ | -0.08 | $(-0.12,0.00)$ | $-0.10(0.095)$ | -0.15 | $(-0.22,0.02)$ |
| 245 | $-0.06(0.294)$ | -0.11 | $(-0.14,0.03)$ | $0.09(0.600)$ | -0.16 | $(-0.66,0.83)$ |
| 246 | $-0.10(0.232)$ | -0.13 | $(-0.16,-0.03)$ | $-0.16(0.137)$ | -0.15 | $(-0.33,0.01)$ |
| 247 | $-0.05(0.292)$ | -0.09 | $(-0.13,0.04)$ | $-0.04(0.168)$ | 0.02 | $(-0.24,0.17)$ |
| 248 | $-0.14(0.135)$ | -0.16 | $(-0.18,-0.10)$ | $-0.19(0.060)$ | -0.19 | $(-0.27,-0.12)$ |
| 249 | $0.00(0.349)$ | -0.04 | $(-0.09,0.10)$ | $0.14(0.523)$ | 0.04 | $(-0.51,0.79)$ |
| 250 | $-0.09(0.219)$ | -0.13 | $(-0.15,-0.03)$ | $-0.07(0.198)$ | -0.02 | $(-0.31,0.18)$ |

Table J-5. Final Discrimination Parameter for $n=250$ Patients

| $\mathbf{A}$ | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | 95\% CI <br> on Mean | Mean (SD) | Median | 95\% CI <br> on Mean |
|  | $0.82(0.226)$ | 0.82 | $(0.77,0.86)$ | $0.83(0.228)$ | 0.84 | $(0.78,0.88)$ |
| 2 | $0.79(0.222)$ | 0.78 | $(0.74,0.83)$ | $0.80(0.222)$ | 0.80 | $(0.75,0.84)$ |
| 3 | $0.77(0.225)$ | 0.78 | $(0.73,0.82)$ | $0.76(0.219)$ | 0.77 | $(0.72,0.81)$ |
| 4 | $0.77(0.220)$ | 0.78 | $(0.73,0.82)$ | $0.77(0.225)$ | 0.77 | $(0.72,0.82)$ |
| 5 | $0.77(0.227)$ | 0.77 | $(0.73,0.82)$ | $0.78(0.238)$ | 0.79 | $(0.73,0.82)$ |
| 6 | $0.77(0.221)$ | 0.76 | $(0.72,0.81)$ | $0.76(0.232)$ | 0.76 | $(0.72,0.81)$ |
| 7 | $0.77(0.222)$ | 0.77 | $(0.72,0.81)$ | $0.77(0.227)$ | 0.76 | $(0.72,0.81)$ |
| 8 | $0.77(0.223)$ | 0.77 | $(0.73,0.82)$ | $0.76(0.209)$ | 0.77 | $(0.72,0.81)$ |
| 9 | $0.77(0.222)$ | 0.77 | $(0.72,0.81)$ | $0.78(0.225)$ | 0.78 | $(0.73,0.83)$ |

Table J-6. Final Difficulty Parameter for $n=250$ Patients

| $\mathbf{y}$ AE | Test |  |  | Control |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Mean (SD) | Median | 95\% CI <br> on Mean | Mean (SD) | Median | 95\% CI <br> on Mean |
|  | $2.82(0.356)$ | 2.80 | $(2.74,2.90)$ | $2.78(0.365)$ | 2.75 | $(2.70,2.86)$ |
| 2 | $3.05(0.403)$ | 3.02 | $(2.96,3.14)$ | $2.96(0.387)$ | 2.93 | $(2.87,3.05)$ |
| 3 | $3.11(0.415)$ | 3.09 | $(3.02,3.20)$ | $3.14(0.442)$ | 3.12 | $(3.05,3.24)$ |
| 4 | $3.18(0.435)$ | 3.15 | $(3.09,3.28)$ | $3.13(0.415)$ | 3.10 | $(3.04,3.22)$ |
| 5 | $3.15(0.427)$ | 3.13 | $(3.06,3.25)$ | $3.19(0.427)$ | 3.16 | $(3.10,3.28)$ |
| 6 | $3.18(0.432)$ | 3.14 | $(3.08,3.27)$ | $3.20(0.453)$ | 3.15 | $(3.10,3.30)$ |
| 7 | $3.20(0.434)$ | 3.16 | $(3.10,3.29)$ | $3.22(0.474)$ | 3.19 | $(3.12,3.33)$ |
| 8 | $3.20(0.441)$ | 3.16 | $(3.11,3.30)$ | $3.19(0.449)$ | 3.15 | $(3.09,3.29)$ |
| 9 | $3.21(0.433)$ | 3.17 | $(3.11,3.30)$ | $3.21(0.444)$ | 3.17 | $(3.12,3.31)$ |

## Appendix K

2-PL EX IRT Model: Superiority and Equivalence Analyses

Table K-1. Superiority Analyses Using Partial Batch Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | $\underset{\text { value) }}{S-W \text { Stat }}(p-$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $\begin{gathered} 9.583 \mathrm{E}-04 \\ (2.636 \mathrm{E}-03) \end{gathered}$ | $(-4.411 \mathrm{E}-03,6.328 \mathrm{E}-03)$ | 64.07 | $\begin{gathered} 3.635 \mathrm{E}-01 \\ (0.719) \end{gathered}$ | $\begin{gathered} 0.5883 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.927 \mathrm{E}-13 \\ (2.862 \mathrm{E}-13) \end{gathered}$ | $(-1.962 \mathrm{E}-15,1.587 \mathrm{E}-12)$ | 97.48 | $\begin{gathered} 2.770 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |
| 2 | Mean | $\begin{aligned} & -1.389 \mathrm{E}-03 \\ & (2.252 \mathrm{E}-03) \end{aligned}$ | (-5.976E-03, 3.198E-03) | 72.91 | $\begin{gathered} -6.168 \mathrm{E}-01 \\ (0.542) \end{gathered}$ | $\begin{gathered} 0.5788 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -1.120 \mathrm{E}-12 \\ & (3.195 \mathrm{E}-13) \end{aligned}$ | $(-2.007 \mathrm{E}-12,-2.329 \mathrm{E}-13)$ | 98.76 | $\begin{gathered} -3.505 \mathrm{E}+00 \\ (0.025) \end{gathered}$ |  |
| 3 | Mean | $\begin{gathered} 1.377 \mathrm{E}-03 \\ (3.047 \mathrm{E}-03) \end{gathered}$ | (-4.829E-03, $7.583 \mathrm{E}-03)$ | 67.28 | $\begin{gathered} 4.520 \mathrm{E}-01 \\ (0.654) \end{gathered}$ | $\begin{gathered} 0.5923 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 8.193 \mathrm{E}-13 \\ (6.837 \mathrm{E}-13) \end{gathered}$ | $(-1.079 \mathrm{E}-12,2.718 \mathrm{E}-12)$ | 85.16 | $\begin{gathered} 1.198 \mathrm{E}+00 \\ (0.297) \end{gathered}$ |  |
| 4 | Mean | $\begin{gathered} -6.193 \mathrm{E}-04 \\ (1.488 \mathrm{E}-03) \end{gathered}$ | $(-3.650 \mathrm{E}-03,2.412 \mathrm{E}-03)$ | 66.00 | $\begin{gathered} -4.162 \mathrm{E}-01 \\ (0.680) \end{gathered}$ | $\begin{gathered} 0.5540 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.889 \mathrm{E}-13 \\ & (3.464 \mathrm{E}-14) \end{aligned}$ | $(-6.850 \mathrm{E}-13,-4.927 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -1.700 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| 5 | Mean | $\begin{gathered} 9.426 \mathrm{E}-04 \\ (2.570 \mathrm{E}-03) \end{gathered}$ | $(-4.293 \mathrm{E}-03,6.178 \mathrm{E}-03)$ | 64.19 | $\begin{gathered} 3.667 \mathrm{E}-01 \\ (0.716) \end{gathered}$ | $\begin{gathered} 0.5891 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.256 \mathrm{E}-13 \\ (3.373 \mathrm{E}-13) \end{gathered}$ | $(-2.108 \mathrm{E}-13,1.662 \mathrm{E}-12)$ | 95.11 | $\begin{gathered} 2.151 \mathrm{E}+00 \\ (0.098) \end{gathered}$ |  |
| 6 | Mean | $\begin{gathered} 1.638 \mathrm{E}-04 \\ (1.393 \mathrm{E}-03) \end{gathered}$ | $(-2.673 \mathrm{E}-03,3.000 \mathrm{E}-03)$ | 54.65 | $\begin{gathered} 1.177 \mathrm{E}-01 \\ (0.907) \end{gathered}$ | $\begin{gathered} 0.5751 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -3.730 \mathrm{E}-14 \\ & (3.753 \mathrm{E}-14) \end{aligned}$ | $(-1.415 \mathrm{E}-13,6.688 \mathrm{E}-14)$ | 81.18 | $\begin{gathered} -9.941 \mathrm{E}-01 \\ (0.376) \end{gathered}$ |  |
| 7 | Mean | $\begin{aligned} & -1.808 \mathrm{E}-04 \\ & (1.403 \mathrm{E}-03) \end{aligned}$ | $(-3.038 \mathrm{E}-03,2.677 \mathrm{E}-03)$ | 55.09 | $\begin{gathered} -1.289 \mathrm{E}-01 \\ (0.898) \end{gathered}$ | $\begin{gathered} 0.5394 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.101 \mathrm{E}-13 \\ (1.710 \mathrm{E}-14) \end{gathered}$ | (6.266E-14, $1.576 \mathrm{E}-13)$ | 99.85 | $\begin{gathered} 6.442 \mathrm{E}+00 \\ (0.003) \end{gathered}$ |  |
| 8 | Mean | $\begin{gathered} 1.684 \mathrm{E}-04 \\ (1.761 \mathrm{E}-03) \end{gathered}$ | $(-3.419 \mathrm{E}-03,3.756 \mathrm{E}-03)$ | 53.78 | $\begin{gathered} 9.561 \mathrm{E}-02 \\ (0.924) \end{gathered}$ | $\begin{gathered} 0.5597 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 2.589 \mathrm{E}-13 \\ (5.773 \mathrm{E}-14) \end{gathered}$ | (9.862E-14, 4.192E-13) | 99.45 | $\begin{gathered} 4.485 \mathrm{E}+00 \\ (0.011) \end{gathered}$ |  |
| 9 | Mean | $\begin{aligned} & -9.689 \mathrm{E}-05 \\ & (1.361 \mathrm{E}-03) \end{aligned}$ | $(-2.869 \mathrm{E}-03,2.675 \mathrm{E}-03)$ | 52.82 | $\begin{gathered} -7.119 \mathrm{E}-02 \\ (0.944) \end{gathered}$ | $\begin{gathered} 0.5536 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (3.553 \mathrm{E}-14) \end{gathered}$ | $(-9.864 \mathrm{E}-14,9.864 \mathrm{E}-14)$ | 50.00 | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (>0.999) \end{gathered}$ |  |
| Overall | Mean | $\begin{gathered} 1.471 \mathrm{E}-04 \\ (1.990 \mathrm{E}-03) \end{gathered}$ | (-3.906E-03, 4.201E-03) | 61.20 | $\begin{gathered} 1.805 \mathrm{E}-02 \\ (0.776) \end{gathered}$ | $\begin{gathered} 0.5700 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.101 \mathrm{E}-13 \\ (5.773 \mathrm{E}-14) \end{gathered}$ | $(-1.415 \mathrm{E}-13,1.576 \mathrm{E}-13)$ | 97.48 | $\begin{gathered} 1.198 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |

Table K-2. Superiority Analyses Using Bootstrap Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $\begin{aligned} & 1.001 \mathrm{E}-03 \\ & (5.726 \mathrm{E}-03) \end{aligned}$ | $(-1.066 \mathrm{E}-02,1.266 \mathrm{E}-02)$ | 77.79 | $\begin{gathered} 4.149 \mathrm{E}-01 \\ (0.444) \end{gathered}$ | $\begin{gathered} 0.5730 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.927 \mathrm{E}-13 \\ (2.862 \mathrm{E}-13) \end{gathered}$ | $(-1.962 \mathrm{E}-15,1.477 \mathrm{E}-12)$ | 97.48 | $\begin{gathered} 2.770 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |
| 2 | Mean | $\begin{aligned} & -1.447 \mathrm{E}-03 \\ & (5.635 \mathrm{E}-03) \end{aligned}$ | $(-1.293 \mathrm{E}-02,1.003 \mathrm{E}-02)$ | 79.89 | $\begin{gathered} -6.174 \mathrm{E}-01 \\ (0.402) \end{gathered}$ | $\begin{gathered} 0.5807 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -1.120 \mathrm{E}-12 \\ & (2.285 \mathrm{E}-13) \end{aligned}$ | $(-1.752 \mathrm{E}-12,-2.259 \mathrm{E}-13)$ | 98.72 | $\begin{gathered} -3.472 \mathrm{E}+00 \\ (0.026) \end{gathered}$ |  |
| 3 | Mean | $\begin{gathered} 1.353 \mathrm{E}-03 \\ (6.263 \mathrm{E}-03) \end{gathered}$ | $(-1.141 \mathrm{E}-02,1.411 \mathrm{E}-02)$ | 76.97 | $\begin{gathered} 4.723 \mathrm{E}-01 \\ (0.461) \end{gathered}$ | $\begin{gathered} 0.5675 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 8.193 \mathrm{E}-13 \\ (5.225 \mathrm{E}-13) \end{gathered}$ | $(-4.480 \mathrm{E}-13,2.452 \mathrm{E}-12)$ | 93.94 | $\begin{gathered} 1.963 \mathrm{E}+00 \\ (0.121) \end{gathered}$ |  |
| 4 | Mean | $\begin{aligned} & -6.324 \mathrm{E}-04 \\ & (4.964 \mathrm{E}-03) \end{aligned}$ | $(-1.074 \mathrm{E}-02,9.478 \mathrm{E}-03)$ | 83.83 | $\begin{gathered} -3.680 \mathrm{E}-01 \\ (0.323) \end{gathered}$ | $\begin{gathered} 0.5585 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.889 \mathrm{E}-13 \\ & (2.642 \mathrm{E}-13) \end{aligned}$ | $(-1.297 \mathrm{E}-12,5.291 \mathrm{E}-14)$ | 95.57 | $\begin{gathered} -2.240 \mathrm{E}+00 \\ (0.089) \end{gathered}$ |  |
| 5 | Mean | $\begin{gathered} 8.923 \mathrm{E}-04 \\ (5.883 \mathrm{E}-03) \end{gathered}$ | $(-1.109 \mathrm{E}-02,1.288 \mathrm{E}-02)$ | 78.74 | $\begin{gathered} 3.945 \mathrm{E}-01 \\ (0.425) \end{gathered}$ | $\begin{gathered} 0.5714 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.256 \mathrm{E}-13 \\ (2.665 \mathrm{E}-13) \end{gathered}$ | $(-5.959 \mathrm{E}-14,1.442 \mathrm{E}-12)$ | 96.16 | $\begin{gathered} 2.370 \mathrm{E}+00 \\ (0.077) \end{gathered}$ |  |
| 6 | Mean | $\begin{gathered} 2.129 \mathrm{E}-04 \\ (4.771 \mathrm{E}-03) \end{gathered}$ | $(-9.505 \mathrm{E}-03,9.930 \mathrm{E}-03)$ | 84.70 | $\begin{gathered} 2.818 \mathrm{E}-01 \\ (0.306) \end{gathered}$ | $\begin{gathered} 0.5727 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -3.730 \mathrm{E}-14 \\ & (3.775 \mathrm{E}-14) \end{aligned}$ | $(-2.123 \mathrm{E}-13,6.688 \mathrm{E}-14)$ | 93.08 | $\begin{gathered} -9.941 \mathrm{E}-01 \\ (0.138) \end{gathered}$ |  |
| 7 | Mean | $\begin{aligned} & -2.151 \mathrm{E}-04 \\ & (4.771 \mathrm{E}-03) \end{aligned}$ | $(-9.932 \mathrm{E}-03,9.502 \mathrm{E}-03)$ | 84.80 | $\begin{gathered} -5.934 \mathrm{E}-02 \\ (0.304) \end{gathered}$ | $\begin{gathered} 0.5429 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.101 \mathrm{E}-13 \\ (5.551 \mathrm{E}-14) \end{gathered}$ | $(-4.522 \mathrm{E}-14,2.643 \mathrm{E}-13)$ | 94.09 | $\begin{gathered} 1.984 \mathrm{E}+00 \\ (0.118) \end{gathered}$ |  |
| 8 | Mean | $\begin{gathered} 2.016 \mathrm{E}-04 \\ (5.255 \mathrm{E}-03) \end{gathered}$ | $(-1.050 \mathrm{E}-02,1.091 \mathrm{E}-02)$ | 81.78 | $\begin{gathered} 1.827 \mathrm{E}-01 \\ (0.364) \end{gathered}$ | $\begin{gathered} 0.5646 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 2.589 \mathrm{E}-13 \\ (1.008 \mathrm{E}-13) \end{gathered}$ | (4.591E-15, $5.187 \mathrm{E}-13)$ | 97.59 | $\begin{gathered} 2.812 \mathrm{E}+00 \\ (0.048) \end{gathered}$ |  |
| 9 | Mean | $\begin{aligned} & -1.328 \mathrm{E}-04 \\ & (4.934 \mathrm{E}-03) \end{aligned}$ | $(-1.018 \mathrm{E}-02,9.917 \mathrm{E}-03)$ | 85.01 | $\begin{gathered} 5.827 \mathrm{E}-02 \\ (0.300) \end{gathered}$ | $\begin{gathered} 0.5589 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (3.553 \mathrm{E}-14) \end{gathered}$ | $(-9.908 \mathrm{E}-14,2.882 \mathrm{E}-14)$ | 50.12 | $\begin{gathered} -1.573 \mathrm{E}-03 \\ (0.998) \end{gathered}$ |  |
| Overall | Mean | $\begin{gathered} 1.371 \mathrm{E}-04 \\ (5.356 \mathrm{E}-03) \end{gathered}$ | $(-1.077 \mathrm{E}-02,1.105 \mathrm{E}-02)$ | 81.50 | $\begin{gathered} 8.440 \mathrm{E}-02 \\ (0.370) \end{gathered}$ | $\begin{gathered} 0.5656 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.101 \mathrm{E}-13 \\ (2.285 \mathrm{E}-13) \end{gathered}$ | (-9.908E-14, 2.643E-13) | 95.57 | $\begin{gathered} 1.963 \mathrm{E}+00 \\ (0.089) \end{gathered}$ |  |

Table K-3. Superiority Analyses Using Jackknife Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (\boldsymbol{p} \text {-value }) \end{gathered}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $\begin{gathered} 9.583 \mathrm{E}-04 \\ (6.844 \mathrm{E}-04) \end{gathered}$ | (-4.357E-04, 2.352E-03) | 91.45 | $\begin{gathered} 1.400 \mathrm{E}+00 \\ (0.171) \end{gathered}$ | $\begin{gathered} 0.5865 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.319 \mathrm{E}-13 \\ (9.437 \mathrm{E}-04) \end{gathered}$ | $(-2.620 \mathrm{E}-03,2.620 \mathrm{E}-03)$ | 50.00 | $\begin{gathered} 7.755 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 2 | Mean | $\begin{aligned} & -1.389 \mathrm{E}-03 \\ & (6.688 \mathrm{E}-04) \end{aligned}$ | $(-2.751 \mathrm{E}-03,-2.655 \mathrm{E}-05)$ | 97.70 | $\begin{gathered} -2.077 \mathrm{E}+00 \\ (0.046) \end{gathered}$ | $\begin{gathered} 0.5767 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -9.355 \mathrm{E}-13 \\ & (1.368 \mathrm{E}-03) \end{aligned}$ | $(-3.797 \mathrm{E}-03,3.797 \mathrm{E}-03)$ | 50.00 | $\begin{gathered} -6.840 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 3 | Mean | $\begin{gathered} 1.377 \mathrm{E}-03 \\ (7.147 \mathrm{E}-04) \end{gathered}$ | $(-7.889 \mathrm{E}-05,2.833 \mathrm{E}-03)$ | 96.85 | $\begin{gathered} 1.927 \mathrm{E}+00 \\ (0.063) \end{gathered}$ | $\begin{gathered} 0.5904 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.487 \mathrm{E}-13 \\ (1.356 \mathrm{E}-03) \end{gathered}$ | $(-3.765 \mathrm{E}-03,3.765 \mathrm{E}-03)$ | 50.00 | $\begin{gathered} 5.522 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 4 | Mean | $\begin{aligned} & -6.193 \mathrm{E}-04 \\ & (6.593 \mathrm{E}-04) \end{aligned}$ | (-1.962E-03, 7.237E-04) | 82.27 | $\begin{gathered} -9.393 \mathrm{E}-01 \\ (0.355) \end{gathered}$ | $\begin{gathered} 0.5519 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.558 \mathrm{E}-13 \\ & (6.098 \mathrm{E}-04) \end{aligned}$ | $(-1.693 \mathrm{E}-03,1.693 \mathrm{E}-03)$ | 50.00 | $\begin{gathered} -9.114 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 5 | Mean | $\begin{gathered} 9.426 \mathrm{E}-04 \\ (6.903 \mathrm{E}-04) \end{gathered}$ | (-4.635E-04, 2.349E-03) | 90.92 | $\begin{gathered} 1.365 \mathrm{E}+00 \\ (0.182) \end{gathered}$ | $\begin{gathered} 0.5872 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 6.701 \mathrm{E}-13 \\ (9.282 \mathrm{E}-04) \end{gathered}$ | $(-2.577 \mathrm{E}-03,2.577 \mathrm{E}-03)$ | 50.00 | $\begin{gathered} 7.220 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 6 | Mean | $\begin{gathered} 1.638 \mathrm{E}-04 \\ (6.668 \mathrm{E}-04) \end{gathered}$ | $(-1.194 \mathrm{E}-03,1.522 \mathrm{E}-03)$ | 59.63 | $\begin{gathered} 2.457 \mathrm{E}-01 \\ (0.807) \end{gathered}$ | $\begin{gathered} 0.5731 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} -5.240 \mathrm{E}-14 \\ (1.613 \mathrm{E}-04) \end{gathered}$ | $(-4.480 \mathrm{E}-04,4.480 \mathrm{E}-04)$ | 50.00 | $\begin{gathered} -3.248 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 7 | Mean | $\begin{aligned} & -1.808 \mathrm{E}-04 \\ & (6.586 \mathrm{E}-04) \end{aligned}$ | $(-1.522 \mathrm{E}-03,1.161 \mathrm{E}-03)$ | 60.73 | $\begin{gathered} -2.745 \mathrm{E}-01 \\ (0.785) \end{gathered}$ | $\begin{gathered} 0.5370 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.097 \mathrm{E}-13 \\ (1.781 \mathrm{E}-04) \end{gathered}$ | (-4.944E-04, 4.944E-04) | 50.00 | $\begin{gathered} 6.161 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 8 | Mean | $\begin{gathered} 1.684 \mathrm{E}-04 \\ (6.634 \mathrm{E}-04) \end{gathered}$ | (-1.183E-03, 1.520E-03) | 59.94 | $\begin{gathered} 2.538 \mathrm{E}-01 \\ (0.801) \end{gathered}$ | $\begin{gathered} 0.5576 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 2.589 \mathrm{E}-13 \\ (1.658 \mathrm{E}-04) \end{gathered}$ | $(-4.604 \mathrm{E}-04,4.604 \mathrm{E}-04)$ | 50.00 | $\begin{gathered} 1.561 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| 9 | Mean | $\begin{aligned} & -9.689 \mathrm{E}-05 \\ & (6.623 \mathrm{E}-04) \end{aligned}$ | $(-1.446 \mathrm{E}-03,1.252 \mathrm{E}-03)$ | 55.77 | $\begin{gathered} -1.463 \mathrm{E}-01 \\ (0.885) \end{gathered}$ | $\begin{gathered} 0.5514 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -2.220 \mathrm{E}-16 \\ & (9.541 \mathrm{E}-05) \end{aligned}$ | $(-2.649 \mathrm{E}-04,2.649 \mathrm{E}-04)$ | 50.00 | $\begin{gathered} -2.327 \mathrm{E}-12 \\ (>0.999) \end{gathered}$ |  |
| Overall | Mean | $\begin{gathered} 1.471 \mathrm{E}-04 \\ (6.743 \mathrm{E}-04) \end{gathered}$ | $(-1.226 \mathrm{E}-03,1.521 \mathrm{E}-03)$ | 77.25 | $\begin{gathered} 1.950 \mathrm{E}-01 \\ (0.455) \end{gathered}$ | $\begin{gathered} 0.5680 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.097 \mathrm{E}-13 \\ (6.098 \mathrm{E}-04) \end{gathered}$ | (-1.693E-03, 1.693E-03) | 50.00 | $\begin{gathered} 5.522 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |

Table K-4. Superiority Analyses Using Partial Batch Estimation to Spline Approximation for $n=30$ Patients

| AE | Estimator | Paired Diff <br> Mean/Med <br> (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $\begin{aligned} & -7.697 \mathrm{E}-03 \\ & (7.615 \mathrm{E}-03) \end{aligned}$ | (-2.321E-02, 7.813E-03) | 84.02 | $\begin{gathered} -1.011 \mathrm{E}+00 \\ (0.320) \end{gathered}$ | $\begin{gathered} 0.4214 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.931 \mathrm{E}-13 \\ (2.864 \mathrm{E}-13) \end{gathered}$ | $(-2.135 \mathrm{E}-15,1.588 \mathrm{E}-12)$ | 97.48 | $\begin{gathered} 2.769 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |
| 2 | Mean | $\begin{aligned} & -1.011 \mathrm{E}-02 \\ & (5.138 \mathrm{E}-03) \end{aligned}$ | $(-2.057 \mathrm{E}-02,3.570 \mathrm{E}-04)$ | 97.11 | $\begin{gathered} -1.967 \mathrm{E}+00 \\ (0.058) \end{gathered}$ | $\begin{gathered} 0.4003 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -7.514 \mathrm{E}-13 \\ & (1.872 \mathrm{E}-13) \end{aligned}$ | (-1.271E-12, -2.317E-13) | 99.20 | $\begin{gathered} -4.014 \mathrm{E}+00 \\ (0.016) \end{gathered}$ |  |
| 3 | Mean | $\begin{aligned} & -7.229 \mathrm{E}-03 \\ & (6.221 \mathrm{E}-03) \end{aligned}$ | (-1.990E-02, 5.443E-03) | 87.31 | $\begin{gathered} -1.162 \mathrm{E}+00 \\ (0.254) \end{gathered}$ | $\begin{gathered} 0.4253 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 8.193 \mathrm{E}-13 \\ (6.839 \mathrm{E}-13) \end{gathered}$ | (-1.079E-12, 2.718E-12) | 85.15 | $\begin{gathered} 1.198 \mathrm{E}+00 \\ (0.297) \end{gathered}$ |  |
| 4 | Mean | $\begin{aligned} & -9.323 \mathrm{E}-03 \\ & (5.383 \mathrm{E}-03) \end{aligned}$ | $(-2.029 \mathrm{E}-02,1.642 \mathrm{E}-03)$ | 95.35 | $\begin{gathered} -1.732 \mathrm{E}+00 \\ (0.093) \end{gathered}$ | $\begin{gathered} 0.4084 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.889 \mathrm{E}-13 \\ & (3.486 \mathrm{E}-14) \end{aligned}$ | (-6.857E-13, -4.921E-13) | >99.99 | $\underset{(<0.001)}{-1.689 \mathrm{E}+01}$ |  |
| 5 | Mean | $\begin{aligned} & -7.669 \mathrm{E}-03 \\ & (5.832 \mathrm{E}-03) \end{aligned}$ | (-1.955E-02, 4.211E-03) | 90.11 | $\begin{gathered} -1.315 \mathrm{E}+00 \\ (0.198) \end{gathered}$ | $\begin{gathered} 0.4229 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.256 \mathrm{E}-13 \\ (3.375 \mathrm{E}-13) \end{gathered}$ | $(-2.114 \mathrm{E}-13,1.663 \mathrm{E}-12)$ | 95.10 | $\begin{gathered} 2.150 \mathrm{E}+00 \\ (0.098) \end{gathered}$ |  |
| 6 | Mean | $\begin{aligned} & -8.502 \mathrm{E}-03 \\ & (5.728 \mathrm{E}-03) \end{aligned}$ | (-2.017E-02, 3.166E-03) | 92.62 | $\begin{gathered} -1.484 \mathrm{E}+00 \\ (0.148) \end{gathered}$ | $\begin{gathered} 0.4159 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -3.730 \mathrm{E}-14 \\ & (3.775 \mathrm{E}-14) \end{aligned}$ | (-1.421E-13, 6.750E-14) | 81.05 | $\begin{gathered} -9.882 \mathrm{E}-01 \\ (0.379) \end{gathered}$ |  |
| 7 | Mean | $\begin{aligned} & -8.838 \mathrm{E}-03 \\ & (5.410 \mathrm{E}-03) \end{aligned}$ | (-1.986E-02, 2.182E-03) | 94.39 | $\begin{gathered} -1.634 \mathrm{E}+00 \\ (0.112) \end{gathered}$ | $\begin{gathered} 0.4134 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.088 \mathrm{E}-13 \\ (3.597 \mathrm{E}-14) \end{gathered}$ | (8.930E-15, 2.087E-13) | 98.05 | $\begin{gathered} 3.025 \mathrm{E}+00 \\ (0.039) \end{gathered}$ |  |
| 8 | Mean | $\begin{aligned} & -8.468 \mathrm{E}-03 \\ & (5.525 \mathrm{E}-03) \end{aligned}$ | (-1.972E-02, 2.786E-03) | 93.24 | $\begin{gathered} -1.533 \mathrm{E}+00 \\ (0.135) \end{gathered}$ | $\begin{gathered} 0.4168 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 2.585 \mathrm{E}-13 \\ (5.795 \mathrm{E}-14) \end{gathered}$ | (9.755E-14, 4.194E-13) | 99.44 | $\begin{gathered} 4.460 \mathrm{E}+00 \\ (0.011) \end{gathered}$ |  |
| 9 | Mean | $\begin{aligned} & -8.750 \mathrm{E}-03 \\ & (5.489 \mathrm{E}-03) \end{aligned}$ | (-1.993E-02, 2.430E-03) | 93.96 | $\begin{gathered} -1.594 \mathrm{E}+00 \\ (0.121) \end{gathered}$ | $\begin{gathered} 0.4144 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (3.553 \mathrm{E}-14) \end{gathered}$ | $(-9.864 \mathrm{E}-14,9.864 \mathrm{E}-14)$ | 50.00 | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (>0.999) \end{gathered}$ |  |
| Overall | Mean | $\begin{aligned} & -8.509 \mathrm{E}-03 \\ & (5.816 \mathrm{E}-03) \end{aligned}$ | $(-2.036 \mathrm{E}-02,3.337 \mathrm{E}-03)$ | 92.01 | $\begin{gathered} -1.492 \mathrm{E}+00 \\ (0.160) \end{gathered}$ | $\begin{gathered} 0.4154 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.088 \mathrm{E}-13 \\ (5.795 \mathrm{E}-14) \end{gathered}$ | (-1.421E-13, 2.087E-13) | 97.48 | $\begin{gathered} 1.198 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |

Table K-5. Superiority Analyses Using Bootstrap Estimation to Spline Approximation for $n=30$ Patients

| AE | Estimator | Paired Diff <br> Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $\begin{aligned} & -7.705 \mathrm{E}-03 \\ & (9.024 \mathrm{E}-03) \end{aligned}$ | (-2.609E-02, 1.068E-02) | 86.59 | $\begin{gathered} -1.353 \mathrm{E}+00 \\ (0.268) \end{gathered}$ | $\begin{gathered} 0.4345 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.931 \mathrm{E}-13 \\ (2.862 \mathrm{E}-13) \end{gathered}$ | $(-2.135 \mathrm{E}-15,1.478 \mathrm{E}-12)$ | 97.48 | $\begin{gathered} 2.769 \mathrm{E}+00 \\ (0.050) \end{gathered}$ |  |
| 2 | Mean | $\begin{aligned} & -1.009 \mathrm{E}-02 \\ & (6.225 \mathrm{E}-03) \end{aligned}$ | $(-2.277 \mathrm{E}-02,2.592 \mathrm{E}-03)$ | 89.99 | $\begin{gathered} -3.767 \mathrm{E}+00 \\ (0.200) \end{gathered}$ | $\begin{gathered} 0.4354 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -7.514 \mathrm{E}-13 \\ & (2.163 \mathrm{E}-13) \end{aligned}$ | $(-1.632 \mathrm{E}-12,-2.317 \mathrm{E}-13)$ | 98.80 | $\begin{gathered} -3.540 \mathrm{E}+00 \\ (0.024) \end{gathered}$ |  |
| 3 | Mean | $\begin{aligned} & -7.203 \mathrm{E}-03 \\ & (7.138 \mathrm{E}-03) \end{aligned}$ | $(-2.174 \mathrm{E}-02,7.337 \mathrm{E}-03)$ | 88.50 | $\begin{gathered} -2.250 \mathrm{E}+00 \\ (0.230) \end{gathered}$ | $\begin{gathered} 0.4342 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 8.193 \mathrm{E}-13 \\ (5.225 \mathrm{E}-13) \end{gathered}$ | $(-4.471 \mathrm{E}-13,2.453 \mathrm{E}-12)$ | 93.91 | $\begin{gathered} 1.958 \mathrm{E}+00 \\ (0.122) \end{gathered}$ |  |
| 4 | Mean | $\begin{aligned} & -9.452 \mathrm{E}-03 \\ & (6.210 \mathrm{E}-03) \end{aligned}$ | $(-2.210 \mathrm{E}-02,3.198 \mathrm{E}-03)$ | 90.05 | $\begin{gathered} -3.743 \mathrm{E}+00 \\ (0.199) \end{gathered}$ | $\begin{gathered} 0.4330 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.889 \mathrm{E}-13 \\ & (2.647 \mathrm{E}-13) \end{aligned}$ | (-1.296E-12, 1.429E-13) | 95.51 | $\begin{gathered} -2.228 \mathrm{E}+00 \\ (0.090) \end{gathered}$ |  |
| 5 | Mean | $\begin{aligned} & -7.366 \mathrm{E}-03 \\ & (6.552 \mathrm{E}-03) \end{aligned}$ | $(-2.071 \mathrm{E}-02,5.980 \mathrm{E}-03)$ | 89.05 | $\begin{gathered} -3.262 \mathrm{E}+00 \\ (0.219) \end{gathered}$ | $\begin{gathered} 0.4357 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.256 \mathrm{E}-13 \\ (2.660 \mathrm{E}-13) \end{gathered}$ | (-6.020E-14, 1.441E-12) | 96.18 | $\begin{gathered} 2.374 \mathrm{E}+00 \\ (0.076) \end{gathered}$ |  |
| 6 | Mean | $\begin{aligned} & -8.148 \mathrm{E}-03 \\ & (6.760 \mathrm{E}-03) \end{aligned}$ | (-2.192E-02, 5.621E-03) | 88.87 | $\begin{gathered} -2.757 \mathrm{E}+00 \\ (0.223) \end{gathered}$ | $\begin{gathered} 0.4354 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -3.730 \mathrm{E}-14 \\ & (3.819 \mathrm{E}-14) \end{aligned}$ | $(-2.143 \mathrm{E}-13,6.750 \mathrm{E}-14)$ | 92.81 | $\begin{gathered} -9.941 \mathrm{E}-01 \\ (0.144) \end{gathered}$ |  |
| 7 | Mean | $\begin{aligned} & -8.666 \mathrm{E}-03 \\ & (6.413 \mathrm{E}-03) \end{aligned}$ | (-2.173E-02, 4.396E-03) | 89.52 | $\begin{gathered} -4.461 \mathrm{E}+00 \\ (0.210) \end{gathered}$ | $\begin{gathered} 0.4363 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.088 \mathrm{E}-13 \\ (3.664 \mathrm{E}-14) \end{gathered}$ | (-4.399E-14, 2.100E-13) | 93.93 | $\begin{gathered} 1.960 \mathrm{E}+00 \\ (0.121) \end{gathered}$ |  |
| 8 | Mean | $\begin{aligned} & -8.512 \mathrm{E}-03 \\ & (6.412 \mathrm{E}-03) \end{aligned}$ | $(-2.157 \mathrm{E}-02,4.548 \mathrm{E}-03)$ | 89.38 | $\begin{gathered} -3.873 \mathrm{E}+00 \\ (0.212) \end{gathered}$ | $\begin{gathered} 0.4352 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 2.585 \mathrm{E}-13 \\ (1.010 \mathrm{E}-13) \end{gathered}$ | (3.974E-15, 5.194E-13) | 97.58 | $\begin{gathered} 2.807 \mathrm{E}+00 \\ (0.048) \end{gathered}$ |  |
| 9 | Mean | $\begin{aligned} & -8.875 \mathrm{E}-03 \\ & (6.349 \mathrm{E}-03) \end{aligned}$ | $(-2.181 \mathrm{E}-02,4.057 \mathrm{E}-03)$ | 89.48 | $\begin{gathered} -3.660 \mathrm{E}+00 \\ (0.210) \end{gathered}$ | $\begin{gathered} 0.4350 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (3.553 \mathrm{E}-14) \end{gathered}$ | $(-9.864 \mathrm{E}-14,3.005 \mathrm{E}-14)$ | 50.00 | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (>0.999) \end{gathered}$ |  |
| Overall | Mean | $\begin{aligned} & -8.446 \mathrm{E}-03 \\ & (6.787 \mathrm{E}-03) \end{aligned}$ | $(-2.227 \mathrm{E}-02,5.379 \mathrm{E}-03)$ | 89.05 | $\begin{gathered} -3.236 \mathrm{E}+00 \\ (0.219) \end{gathered}$ | $\begin{gathered} 0.4350 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 1.088 \mathrm{E}-13 \\ (2.163 \mathrm{E}-13) \end{gathered}$ | $(-9.864 \mathrm{E}-14,2.100 \mathrm{E}-13)$ | 95.51 | $\begin{gathered} 1.958 \mathrm{E}+00 \\ (0.090) \end{gathered}$ |  |

Table K-6. Superiority Analyses Using Jackknife Estimation to Spline Approximation for $n=30$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p-\text { value }) \end{gathered}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $\begin{aligned} & -7.697 \mathrm{E}-03 \\ & (3.133 \mathrm{E}-03) \end{aligned}$ | (-1.408E-02, -1.316E-03) | 99.02 | $\begin{gathered} -2.457 \mathrm{E}+00 \\ (0.020) \end{gathered}$ | $\begin{gathered} 0.4216 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.321 \mathrm{E}-13 \\ (7.580 \mathrm{E}-03) \end{gathered}$ | $(-2.105 \mathrm{E}-02,2.105 \mathrm{E}-02)$ | 50.00 | $\begin{gathered} 9.658 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| 2 | Mean | $\begin{aligned} & -1.011 \mathrm{E}-02 \\ & (3.016 \mathrm{E}-03) \end{aligned}$ | (-1.625E-02, -3.964E-03) | 99.90 | $\begin{gathered} -3.351 \mathrm{E}+00 \\ (0.002) \end{gathered}$ | $\begin{gathered} 0.4032 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -9.330 \mathrm{E}-13 \\ & (9.954 \mathrm{E}-03) \end{aligned}$ | $(-2.764 \mathrm{E}-02,2.764 \mathrm{E}-02)$ | 50.00 | $\underset{(>0.999)}{-9.374 \mathrm{E}-11}$ |  |
| 3 | Mean | $\begin{aligned} & -7.229 \mathrm{E}-03 \\ & (3.159 \mathrm{E}-03) \end{aligned}$ | (-1.366E-02, -7.955E-04) | 98.56 | $\begin{gathered} -2.289 \mathrm{E}+00 \\ (0.029) \end{gathered}$ | $\begin{gathered} 0.4251 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 7.492 \mathrm{E}-13 \\ (7.119 \mathrm{E}-03) \end{gathered}$ | (-1.977E-02, 1.977E-02) | 50.00 | $\underset{(>0.999)}{1.052 \mathrm{E}-10}$ |  |
| 4 | Mean | $\begin{aligned} & -9.323 \mathrm{E}-03 \\ & (3.069 \mathrm{E}-03) \end{aligned}$ | (-1.558E-02, -3.071E-03) | 99.76 | $\begin{gathered} -3.038 \mathrm{E}+00 \\ (0.005) \end{gathered}$ | $\begin{gathered} 0.4102 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.560 \mathrm{E}-13 \\ & (9.181 \mathrm{E}-03) \end{aligned}$ | $(-2.549 \mathrm{E}-02,2.549 \mathrm{E}-02)$ | 50.00 | $\underset{(>0.999)}{-6.056 \mathrm{E}-11}$ |  |
| 5 | Mean | $\begin{aligned} & -7.669 \mathrm{E}-03 \\ & (3.116 \mathrm{E}-03) \end{aligned}$ | (-1.402E-02, -1.322E-03) | 99.03 | $\begin{gathered} -2.461 \mathrm{E}+00 \\ (0.019) \end{gathered}$ | $\begin{gathered} 0.4231 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 6.699 \mathrm{E}-13 \\ (7.552 \mathrm{E}-03) \end{gathered}$ | $(-2.097 \mathrm{E}-02,2.097 \mathrm{E}-02)$ | 50.00 | $\underset{\substack{8.871 \mathrm{E}-11 \\(>0.999)}}{ }$ |  |
| 6 | Mean | $\begin{aligned} & -8.502 \mathrm{E}-03 \\ & (3.093 \mathrm{E}-03) \end{aligned}$ | (-1.480E-02, -2.201E-03) | 99.51 | $\begin{gathered} -2.748 \mathrm{E}+00 \\ (0.010) \end{gathered}$ | $\begin{gathered} 0.4169 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{aligned} & -5.218 \mathrm{E}-14 \\ & (8.372 \mathrm{E}-03) \end{aligned}$ | $(-2.324 \mathrm{E}-02,2.324 \mathrm{E}-02)$ | 50.00 | $\begin{gathered} -6.233 \mathrm{E}-12 \\ (>0.999) \end{gathered}$ |  |
| 7 | Mean | $\begin{aligned} & -8.838 \mathrm{E}-03 \\ & (3.043 \mathrm{E}-03) \end{aligned}$ | (-1.504E-02, -2.639E-03) | 99.67 | $\begin{gathered} -2.904 \mathrm{E}+00 \\ (0.007) \end{gathered}$ | $\begin{gathered} 0.4149 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 9.348 \mathrm{E}-14 \\ (8.703 \mathrm{E}-03) \end{gathered}$ | $(-2.416 \mathrm{E}-02,2.416 \mathrm{E}-02)$ | 50.00 | $\begin{gathered} 1.074 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| 8 | Mean | $\begin{aligned} & -8.468 \mathrm{E}-03 \\ & (3.059 \mathrm{E}-03) \end{aligned}$ | (-1.470E-02, -2.238E-03) | 99.54 | $\begin{gathered} -2.769 \mathrm{E}+00 \\ (0.009) \end{gathered}$ | $\begin{gathered} 0.4178 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 2.587 \mathrm{E}-13 \\ (8.339 \mathrm{E}-03) \end{gathered}$ | (-2.315E-02, 2.315E-02) | 50.00 | $\underset{(>0.999)}{3.102 \mathrm{E}-11}$ |  |
| 9 | Mean | $\begin{aligned} & -8.750 \mathrm{E}-03 \\ & (3.060 \mathrm{E}-03) \end{aligned}$ | (-1.498E-02, -2.518E-03) | 99.63 | $\begin{gathered} -2.860 \mathrm{E}+00 \\ (0.007) \end{gathered}$ | $\begin{gathered} 0.4157 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (8.617 \mathrm{E}-03) \end{gathered}$ | $(-2.392 \mathrm{E}-02,2.392 \mathrm{E}-02)$ | 50.00 | $\begin{gathered} 0.000 \mathrm{E}+00 \\ (>0.999) \end{gathered}$ |  |
| Overall | Mean | $\begin{aligned} & -8.509 \mathrm{E}-03 \\ & (3.083 \mathrm{E}-03) \end{aligned}$ | (-1.479E-02, -2.229E-03) | 99.40 | $\begin{gathered} -2.764 \mathrm{E}+00 \\ (0.012) \end{gathered}$ | $\begin{gathered} 0.4165 \\ (<0.0001) \end{gathered}$ |
|  | Median | $\begin{gathered} 9.348 \mathrm{E}-14 \\ (8.372 \mathrm{E}-03) \end{gathered}$ | $(-2.324 \mathrm{E}-02,2.324 \mathrm{E}-02)$ | 50.00 | $\underset{(>0.999)}{1.074 \mathrm{E}-11}$ |  |

Table K-7. Superiority Analyses Using Partial Batch Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat ( $p$-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $-9.372 \mathrm{E}-05(2.256 \mathrm{E}-03)$ | (-4.537E-03, 4.349E-03) | 51.66 | $\begin{gathered} -4.155 \mathrm{E}-02 \\ (0.967) \end{gathered}$ | $\begin{gathered} 0.1978 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-9.592 \mathrm{E}-14(3.131 \mathrm{E}-14)$ | $(-1.631 \mathrm{E}-13,-2.877 \mathrm{E}-14)$ | 99.58 | $\begin{gathered} -3.064 \mathrm{E}+00 \\ (0.008) \end{gathered}$ |  |
| 2 | Mean | $4.127 \mathrm{E}-06$ (2.089E-03) | $(-4.109 \mathrm{E}-03,4.117 \mathrm{E}-03)$ | 50.08 | $\begin{gathered} 1.976 \mathrm{E}-03 \\ (0.998) \end{gathered}$ | $\begin{gathered} 0.1855 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-13$ (2.998E-14) | (1.089E-13, $2.375 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 5.778 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 3 | Mean | $-1.486 \mathrm{E}-04(2.177 \mathrm{E}-03)$ | (-4.437E-03, 4.139E-03) | 52.72 | $\begin{gathered} -6.827 \mathrm{E}-02 \\ (0.946) \end{gathered}$ | $\begin{gathered} 0.2086 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.130 \mathrm{E}-14(6.661 \mathrm{E}-16)$ | $(-4.273 \mathrm{E}-14,-3.987 \mathrm{E}-14)$ | >99.99 | $\begin{gathered} -6.200 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| 4 | Mean | $-3.705 \mathrm{E}-05(2.052 \mathrm{E}-03)$ | $(-4.079 \mathrm{E}-03,4.005 \mathrm{E}-03)$ | 50.72 | $\begin{gathered} -1.805 \mathrm{E}-02 \\ (0.986) \end{gathered}$ | $\begin{gathered} 0.1918 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.594 \mathrm{E}-13(2.753 \mathrm{E}-14)$ | (1.004E-13, $2.185 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 5.790 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 5 | Mean | $-2.359 \mathrm{E}-04(1.938 \mathrm{E}-03)$ | $(-4.052 \mathrm{E}-03,3.580 \mathrm{E}-03)$ | 54.84 | $\begin{gathered} -1.217 \mathrm{E}-01 \\ (0.903) \end{gathered}$ | $\begin{gathered} 0.2185 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-5.334 \mathrm{E}-13(1.554 \mathrm{E}-15)$ | $(-5.367 \mathrm{E}-13,-5.300 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -3.431 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ |  |
| 6 | Mean | $-1.281 \mathrm{E}-04(2.160 \mathrm{E}-03)$ | $(-4.383 \mathrm{E}-03,4.126 \mathrm{E}-03)$ | 52.36 | $\begin{gathered} -5.931 \mathrm{E}-02 \\ (0.953) \end{gathered}$ | $\begin{gathered} 0.2055 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-14(6.661 \mathrm{E}-16)$ | (1.589E-14, $1.875 \mathrm{E}-14)$ | >99.99 | $\begin{gathered} 2.600 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| 7 | Mean | $-1.633 \mathrm{E}-04(2.040 \mathrm{E}-03)$ | $(-4.181 \mathrm{E}-03,3.855 \mathrm{E}-03)$ | 53.19 | $\begin{gathered} -8.005 \mathrm{E}-02 \\ (0.936) \\ \hline \end{gathered}$ | $\begin{gathered} 0.2089 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-2.047 \mathrm{E}-13$ (8.882E-16) | $(-2.066 \mathrm{E}-13,-2.028 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -2.305 \mathrm{E}+02 \\ (<0.001) \end{gathered}$ |  |
| 8 | Mean | $-7.377 \mathrm{E}-05(2.154 \mathrm{E}-03)$ | $(-4.316 \mathrm{E}-03,4.169 \mathrm{E}-03)$ | 51.36 | $\begin{gathered} -3.425 \mathrm{E}-02 \\ (0.973) \end{gathered}$ | $\begin{gathered} 0.1980 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-13$ (2.665E-14) | (1.160E-13, $2.303 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 6.500 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 9 | Mean | $-1.623 \mathrm{E}-04(1.922 \mathrm{E}-03)$ | $(-3.948 \mathrm{E}-03,3.623 \mathrm{E}-03)$ | 53.36 | $\begin{gathered} -8.442 \mathrm{E}-02 \\ (0.933) \end{gathered}$ | $\begin{gathered} 0.2070 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-3.193 \mathrm{E}-13$ (2.465E-14) | $(-3.722 \mathrm{E}-13,-2.664 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -1.295 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| Overall | Mean | $-1.154 \mathrm{E}-04(2.088 \mathrm{E}-03)$ | (-4.227E-03, 3.996E-03) | 52.25 | $\begin{gathered} -5.618 \mathrm{E}-02 \\ (0.955) \end{gathered}$ | $\begin{gathered} 0.2024 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.130 \mathrm{E}-14(2.465 \mathrm{E}-14)$ | $(-4.273 \mathrm{E}-14,-2.877 \mathrm{E}-14)$ | >99.99 | $\begin{gathered} -3.064 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |

Table K-8. Superiority Analyses Using Bootstrap Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $-8.031 \mathrm{E}-05(3.270 \mathrm{E}-03)$ | (-6.520E-03, 6.359E-03) | 61.87 | $\begin{gathered} 2.946 \mathrm{E}-02 \\ (0.763) \end{gathered}$ | $\begin{gathered} 0.2731 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-9.592 \mathrm{E}-14$ (3.131E-14) | $(-1.626 \mathrm{E}-13,-2.877 \mathrm{E}-14)$ | 99.58 | $\begin{gathered} -3.064 \mathrm{E}+00 \\ (0.008) \end{gathered}$ |  |
| 2 | Mean | $-1.027 \mathrm{E}-05(3.093 \mathrm{E}-03)$ | (-6.102E-03, 6.082E-03) | 63.11 | $\begin{gathered} 6.753 \mathrm{E}-02 \\ (0.738) \end{gathered}$ | $\begin{gathered} 0.2643 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-13$ (2.975E-14) | (1.127E-13, $2.375 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 6.077 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 3 | Mean | $-1.515 \mathrm{E}-04(3.148 \mathrm{E}-03)$ | (-6.351E-03, 6.048E-03) | 61.78 | $\begin{gathered} -9.953 \mathrm{E}-04 \\ (0.764) \end{gathered}$ | $\begin{gathered} 0.2790 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.130 \mathrm{E}-14(2.687 \mathrm{E}-14)$ | $(-9.893 \mathrm{E}-14,1.544 \mathrm{E}-14)$ | 92.98 | $\begin{gathered} -1.563 \mathrm{E}+00 \\ (0.140) \end{gathered}$ |  |
| 4 | Mean | $-3.438 \mathrm{E}-05(3.033 \mathrm{E}-03)$ | $(-6.007 \mathrm{E}-03,5.938 \mathrm{E}-03)$ | 63.28 | $\begin{gathered} 5.802 \mathrm{E}-02 \\ (0.734) \end{gathered}$ | $\begin{gathered} 0.2682 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.594 \mathrm{E}-13$ (2.731E-14) | (1.014E-13, $2.185 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 5.917 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 5 | Mean | $-2.359 \mathrm{E}-04(2.954 \mathrm{E}-03)$ | $(-6.053 \mathrm{E}-03,5.581 \mathrm{E}-03)$ | 62.80 | $\begin{gathered} -4.086 \mathrm{E}-02 \\ (0.744) \end{gathered}$ | $\begin{gathered} 0.2895 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-5.334 \mathrm{E}-13(2.442 \mathrm{E}-14)$ | $(-5.857 \mathrm{E}-13,-4.822 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -2.174 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| 6 | Mean | $-1.108 \mathrm{E}-04(3.120 \mathrm{E}-03)$ | (-6.256E-03, 6.034E-03) | 62.07 | $\begin{gathered} 1.742 \mathrm{E}-02 \\ (0.759) \end{gathered}$ | $\begin{gathered} 0.2750 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-14(2.598 \mathrm{E}-14)$ | $(-3.840 \mathrm{E}-14,7.215 \mathrm{E}-14)$ | 74.57 | $\begin{gathered} 6.667 \mathrm{E}-01 \\ (0.509) \end{gathered}$ |  |
| 7 | Mean | $-1.387 \mathrm{E}-04(3.026 \mathrm{E}-03)$ | $(-6.098 \mathrm{E}-03,5.821 \mathrm{E}-03)$ | 62.64 | $\begin{gathered} 7.048 \mathrm{E}-03 \\ (0.747) \end{gathered}$ | $\begin{gathered} 0.2795 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-2.047 \mathrm{E}-13(2.509 \mathrm{E}-14)$ | $(-2.585 \mathrm{E}-13,-1.519 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -8.212 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 8 | Mean | $-6.939 \mathrm{E}-05(3.082 \mathrm{E}-03)$ | $(-6.138 \mathrm{E}-03,6.000 \mathrm{E}-03)$ | 62.41 | $\begin{gathered} 3.629 \mathrm{E}-02 \\ (0.752) \\ \hline \end{gathered}$ | $\begin{gathered} 0.2693 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-13$ (2.665E-14) | (1.175E-13, 2.303E-13) | >99.99 | $\begin{gathered} 6.667 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 9 | Mean | $-1.492 \mathrm{E}-04(2.930 \mathrm{E}-03)$ | $(-5.919 \mathrm{E}-03,5.621 \mathrm{E}-03)$ | 63.23 | $\begin{gathered} 4.804 \mathrm{E}-03 \\ (0.735) \end{gathered}$ | $\begin{gathered} 0.2794 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-3.193 \mathrm{E}-13$ (2.442E-14) | $(-3.717 \mathrm{E}-13,-2.669 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -1.295 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| Overall | Mean | $-1.089 \mathrm{E}-04(3.073 \mathrm{E}-03)$ | $(-6.161 \mathrm{E}-03,5.943 \mathrm{E}-03)$ | 62.58 | $\begin{gathered} 1.986 \mathrm{E}-02 \\ (0.748) \end{gathered}$ | $\begin{gathered} 0.2753 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.130 \mathrm{E}-14(2.665 \mathrm{E}-14)$ | $(-9.893 \mathrm{E}-14,1.544 \mathrm{E}-14)$ | >99.99 | $\begin{gathered} -1.563 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |

Table K-9. Superiority Analyses Using Jackknife Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p-\text { value }) \end{gathered}$ | S-W Stat ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | -9.372E-05 (6.629E-05) | (-2.243E-04, 3.683E-05) | 92.07 | $\begin{gathered} -1.414 \mathrm{E}+00 \\ (0.159) \end{gathered}$ | $\begin{gathered} 0.1984 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-9.592 \mathrm{E}-14(9.354 \mathrm{E}-05)$ | $(-2.006 \mathrm{E}-04,2.006 \mathrm{E}-04)$ | 50.00 | $\begin{gathered} -1.026 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| 2 | Mean | $4.127 \mathrm{E}-06$ (7.136E-05) | (-1.364E-04, 1.447E-04) | 52.30 | $\begin{gathered} 5.784 \mathrm{E}-02 \\ (0.954) \end{gathered}$ | $\begin{gathered} 0.1861 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.743 \mathrm{E}-13$ (4.119E-06) | $(-8.834 \mathrm{E}-06,8.834 \mathrm{E}-06)$ | 50.00 | $\begin{gathered} 4.232 \mathrm{E}-08 \\ (>0.999) \end{gathered}$ |  |
| 3 | Mean | $-1.486 \mathrm{E}-04(6.357 \mathrm{E}-05)$ | $(-2.738 \mathrm{E}-04,-2.343 \mathrm{E}-05)$ | 98.99 | $\begin{gathered} -2.338 \mathrm{E}+00 \\ (0.020) \end{gathered}$ | $\begin{gathered} 0.2091 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.130 \mathrm{E}-14(1.483 \mathrm{E}-04)$ | (-3.182E-04, 3.182E-04) | 50.00 | $\begin{gathered} -2.784 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 4 | Mean | $-3.705 \mathrm{E}-05(6.932 \mathrm{E}-05)$ | (-1.736E-04, 9.947E-05) | 70.32 | $\begin{gathered} -5.344 \mathrm{E}-01 \\ (0.594) \end{gathered}$ | $\begin{gathered} 0.1923 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.594 \mathrm{E}-13$ (3.697E-05) | (-7.930E-05, 7.930E-05) | 50.00 | $\begin{gathered} 4.312 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| 5 | Mean | $-2.359 \mathrm{E}-04(6.060 \mathrm{E}-05)$ | (-3.552E-04, -1.165E-04) | 99.99 | $\begin{gathered} -3.892 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.2191 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-5.334 \mathrm{E}-13(2.354 \mathrm{E}-04)$ | $(-5.049 \mathrm{E}-04,5.049 \mathrm{E}-04)$ | 50.00 | $\begin{gathered} -2.266 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| 6 | Mean | $-1.281 \mathrm{E}-04(6.482 \mathrm{E}-05)$ | $(-2.558 \mathrm{E}-04,-4.728 \mathrm{E}-07)$ | 97.54 | $\begin{gathered} -1.977 \mathrm{E}+00 \\ (0.049) \end{gathered}$ | $\begin{gathered} 0.2060 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.732 \mathrm{E}-14$ (1.279E-04) | (-2.743E-04, 2.743E-04) | 50.00 | $\begin{gathered} 1.354 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 7 | Mean | $-1.633 \mathrm{E}-04(6.370 \mathrm{E}-05)$ | $(-2.888 \mathrm{E}-04,-3.789 \mathrm{E}-05)$ | 99.45 | $\begin{gathered} -2.564 \mathrm{E}+00 \\ (0.011) \end{gathered}$ | $\begin{gathered} 0.2094 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-2.047 \mathrm{E}-13$ (1.630E-04) | (-3.496E-04, 3.496E-04) | 50.00 | $\begin{gathered} -1.256 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| 8 | Mean | $-7.377 \mathrm{E}-05(6.735 \mathrm{E}-05)$ | $(-2.064 \mathrm{E}-04,5.887 \mathrm{E}-05)$ | 86.28 | $\begin{gathered} -1.095 \mathrm{E}+00 \\ (0.274) \end{gathered}$ | $\begin{gathered} 0.1985 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.734 \mathrm{E}-13$ (7.363E-05) | (-1.579E-04, 1.579E-04) | 50.00 | $\begin{gathered} 2.355 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| 9 | Mean | $-1.623 \mathrm{E}-04(6.416 \mathrm{E}-05)$ | (-2.886E-04, -3.590E-05) | 99.40 | $\begin{gathered} -2.529 \mathrm{E}+00 \\ (0.012) \end{gathered}$ | $\begin{gathered} 0.2075 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-3.193 \mathrm{E}-13$ (1.619E-04) | (-3.473E-04, 3.473E-04) | 50.00 | $\begin{gathered} -1.972 \mathrm{E}-09 \\ (>0.999) \end{gathered}$ |  |
| Overall | Mean | $-1.154 \mathrm{E}-04(6.568 \mathrm{E}-05)$ | (-2.448E-04, 1.396E-05) | 88.48 | $\begin{gathered} -1.810 \mathrm{E}+00 \\ (0.230) \end{gathered}$ | $\begin{gathered} 0.2029 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.130 \mathrm{E}-14(1.279 \mathrm{E}-04)$ | $(-2.743 \mathrm{E}-04,2.743 \mathrm{E}-04)$ | 50.00 | $\begin{gathered} -2.784 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |

Table K-10. Superiority Analyses Using Partial Batch Estimation to Spline Approximation for $n=250$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $-3.501 \mathrm{E}-03(1.428 \mathrm{E}-02)$ | (-3.163E-02, 2.463E-02) | 59.67 | $\begin{gathered} -2.451 \mathrm{E}-01 \\ (0.807) \end{gathered}$ | $\begin{gathered} 0.0876 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-9.592 \mathrm{E}-14$ (3.153E-14) | $(-1.635 \mathrm{E}-13,-2.830 \mathrm{E}-14)$ | 99.56 | $\begin{gathered} -3.042 \mathrm{E}+00 \\ (0.009) \end{gathered}$ |  |
| 2 | Mean | $-3.389 \mathrm{E}-03$ (1.412E-02) | $(-3.120 \mathrm{E}-02,2.443 \mathrm{E}-02)$ | 59.47 | $\begin{gathered} -2.400 \mathrm{E}-01 \\ (0.811) \end{gathered}$ | $\begin{gathered} 0.0882 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.741 \mathrm{E}-13$ (2.975E-14) | (1.103E-13, $2.379 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 5.851 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 3 | Mean | $-3.540 \mathrm{E}-03(1.397 \mathrm{E}-02)$ | (-3.106E-02, 2.398E-02) | 59.99 | $\begin{gathered} -2.533 \mathrm{E}-01 \\ (0.800) \end{gathered}$ | $\begin{gathered} 0.0878 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.174 \mathrm{E}-14(2.665 \mathrm{E}-14)$ | $(-9.889 \mathrm{E}-14,1.540 \mathrm{E}-14)$ | 93.02 | $\begin{gathered} -1.567 \mathrm{E}+00 \\ (0.140) \end{gathered}$ |  |
| 4 | Mean | $-3.423 \mathrm{E}-03(1.399 \mathrm{E}-02)$ | (-3.097E-02, 2.412E-02) | 59.66 | $\begin{gathered} -2.447 \mathrm{E}-01 \\ (0.807) \end{gathered}$ | $\begin{gathered} 0.0882 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.590 \mathrm{E}-13$ (2.753E-14) | (9.993E-14, 2.180E-13) | >99.99 | $\begin{gathered} 5.774 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 5 | Mean | $-3.621 \mathrm{E}-03(1.374 \mathrm{E}-02)$ | $(-3.069 \mathrm{E}-02,2.345 \mathrm{E}-02)$ | 60.38 | $\begin{gathered} -2.634 \mathrm{E}-01 \\ (0.792) \end{gathered}$ | $\begin{gathered} 0.0877 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-5.338 \mathrm{E}-13$ (2.531E-14) | $(-5.881 \mathrm{E}-13,-4.795 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -2.109 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| 6 | Mean | $-3.513 \mathrm{E}-03(1.389 \mathrm{E}-02)$ | $(-3.086 \mathrm{E}-02,2.384 \mathrm{E}-02)$ | 59.97 | $\begin{gathered} -2.530 \mathrm{E}-01 \\ (0.801) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.688 \mathrm{E}-14(2.576 \mathrm{E}-14)$ | $(-3.837 \mathrm{E}-14,7.212 \mathrm{E}-14)$ | 73.85 | $\begin{gathered} 6.552 \mathrm{E}-01 \\ (0.523) \end{gathered}$ |  |
| 7 | Mean | $-3.545 \mathrm{E}-03(1.378 \mathrm{E}-02)$ | $(-3.069 \mathrm{E}-02,2.360 \mathrm{E}-02)$ | 60.14 | $\begin{gathered} -2.572 \mathrm{E}-01 \\ (0.797) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-2.043 \mathrm{E}-13(2.487 \mathrm{E}-14)$ | $(-2.576 \mathrm{E}-13,-1.509 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -8.214 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 8 | Mean | $-3.457 \mathrm{E}-03$ (1.393E-02) | (-3.089E-02, 2.397E-02) | 59.79 | $\begin{gathered} -2.482 \mathrm{E}-01 \\ (0.804) \\ \hline \end{gathered}$ | $\begin{gathered} 0.0882 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.741 \mathrm{E}-13$ (2.620E-14) | (1.179E-13, 2.303E-13) | >99.99 | $\begin{gathered} 6.644 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 9 | Mean | $-3.543 \mathrm{E}-03(1.375 \mathrm{E}-02)$ | $(-3.061 \mathrm{E}-02,2.353 \mathrm{E}-02)$ | 60.16 | $\begin{gathered} -2.577 \mathrm{E}-01 \\ (0.797) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-3.189 \mathrm{E}-13$ (2.431E-14) | $(-3.710 \mathrm{E}-13,-2.667 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -1.311 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| Overall | Mean | $-3.504 \mathrm{E}-03(1.394 \mathrm{E}-02)$ | (-3.096E-02, 2.395E-02) | 59.91 | $\begin{gathered} -2.514 \mathrm{E}-01 \\ (0.802) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.174 \mathrm{E}-14(2.620 \mathrm{E}-14)$ | $(-9.889 \mathrm{E}-14,1.540 \mathrm{E}-14)$ | >99.99 | $\begin{gathered} -1.567 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |

Table K-11. Superiority Analyses Using Bootstrap Estimation to Spline Approximation for $n=250$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $-3.624 \mathrm{E}-03$ (1.105E-02) | $(-2.539 \mathrm{E}-02,1.814 \mathrm{E}-02)$ | 77.24 | $\begin{gathered} -9.775 \mathrm{E}-01 \\ (0.455) \end{gathered}$ | $\begin{gathered} 0.1477 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-9.592 \mathrm{E}-14$ (3.131E-14) | $(-1.630 \mathrm{E}-13,-2.925 \mathrm{E}-14)$ | 99.56 | $\begin{gathered} -3.042 \mathrm{E}+00 \\ (0.009) \end{gathered}$ |  |
| 2 | Mean | $-3.425 \mathrm{E}-03(1.111 \mathrm{E}-02)$ | $(-2.530 \mathrm{E}-02,1.845 \mathrm{E}-02)$ | 76.81 | $\begin{gathered} -7.462 \mathrm{E}-01 \\ (0.464) \end{gathered}$ | $\begin{gathered} 0.1481 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.741 \mathrm{E}-13(2.931 \mathrm{E}-14)$ | (1.131E-13, $2.379 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 6.125 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 3 | Mean | $-3.462 \mathrm{E}-03$ (1.155E-02) | $(-2.621 \mathrm{E}-02,1.929 \mathrm{E}-02)$ | 75.36 | $\begin{gathered} -6.046 \mathrm{E}-01 \\ (0.493) \end{gathered}$ | $\begin{gathered} 0.1503 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.174 \mathrm{E}-14(2.665 \mathrm{E}-14)$ | $(-9.913 \mathrm{E}-14,1.540 \mathrm{E}-14)$ | 93.02 | $\begin{gathered} -1.567 \mathrm{E}+00 \\ (0.140) \end{gathered}$ |  |
| 4 | Mean | $-3.421 \mathrm{E}-03(1.139 \mathrm{E}-02)$ | $(-2.585 \mathrm{E}-02,1.901 \mathrm{E}-02)$ | 75.69 | $\begin{gathered} -6.187 \mathrm{E}-01 \\ (0.486) \end{gathered}$ | $\begin{gathered} 0.1484 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.590 \mathrm{E}-13$ (2.709E-14) | (1.018E-13, $2.180 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} 5.967 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 5 | Mean | -3.585E-03 (1.125E-02) | $(-2.573 \mathrm{E}-02,1.856 \mathrm{E}-02)$ | 75.88 | $\begin{gathered} -6.659 \mathrm{E}-01 \\ (0.482) \end{gathered}$ | $\begin{gathered} 0.1493 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-5.338 \mathrm{E}-13$ (2.442E-14) | $(-5.881 \mathrm{E}-13,-4.822 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -2.185 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| 6 | Mean | $-3.688 \mathrm{E}-03$ (1.146E-02) | $(-2.626 \mathrm{E}-02,1.888 \mathrm{E}-02)$ | 75.77 | $\begin{gathered} -6.554 \mathrm{E}-01 \\ (0.485) \end{gathered}$ | $\begin{gathered} 0.1493 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.688 \mathrm{E}-14$ (2.576E-14) | $(-3.837 \mathrm{E}-14,7.212 \mathrm{E}-14)$ | 73.85 | $\begin{gathered} 6.552 \mathrm{E}-01 \\ (0.523) \end{gathered}$ |  |
| 7 | Mean | $-3.495 \mathrm{E}-03$ (1.126E-02) | $(-2.567 \mathrm{E}-02,1.868 \mathrm{E}-02)$ | 75.80 | $\begin{gathered} -6.228 \mathrm{E}-01 \\ (0.484) \end{gathered}$ | $\begin{gathered} 0.1482 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-2.043 \mathrm{E}-13(2.487 \mathrm{E}-14)$ | $(-2.576 \mathrm{E}-13,-1.527 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -8.286 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 8 | Mean | $-3.285 \mathrm{E}-03(1.145 \mathrm{E}-02)$ | $(-2.583 \mathrm{E}-02,1.926 \mathrm{E}-02)$ | 75.43 | $\begin{gathered} -5.751 \mathrm{E}-01 \\ (0.491) \end{gathered}$ | $\begin{gathered} 0.1492 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.741 \mathrm{E}-13$ (2.620E-14) | (1.179E-13, 2.303E-13) | >99.99 | $\begin{gathered} 6.644 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |
| 9 | Mean | $-3.496 \mathrm{E}-03(1.141 \mathrm{E}-02)$ | $(-2.597 \mathrm{E}-02,1.898 \mathrm{E}-02)$ | 75.71 | $\begin{gathered} -6.408 \mathrm{E}-01 \\ (0.486) \end{gathered}$ | $\begin{gathered} 0.1517 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-3.189 \mathrm{E}-13$ (2.442E-14) | $(-3.712 \mathrm{E}-13,-2.673 \mathrm{E}-13)$ | >99.99 | $\begin{gathered} -1.305 \mathrm{E}+01 \\ (<0.001) \end{gathered}$ |  |
| Overall | Mean | $-3.498 \mathrm{E}-03(1.133 \mathrm{E}-02)$ | $(-2.580 \mathrm{E}-02,1.881 \mathrm{E}-02)$ | 75.97 | $\begin{gathered} -6.786 \mathrm{E}-01 \\ (0.481) \end{gathered}$ | $\begin{gathered} 0.1491 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.174 \mathrm{E}-14(2.620 \mathrm{E}-14)$ | $(-9.913 \mathrm{E}-14,1.540 \mathrm{E}-14)$ | >99.99 | $\begin{gathered} -1.567 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ |  |

Table K-12. Superiority Analyses Using Jackknife Estimation to Spline Approximation for $n=250$ Patients

| AE | Estimator | Paired Diff Mean/Med (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Mean | $-3.501 \mathrm{E}-03$ (4.802E-04) | (-4.447E-03, -2.555E-03) | >99.99 | $\underset{(<0.001)}{-7.291 \mathrm{E}+00}$ | $\begin{gathered} 0.0876 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-9.592 \mathrm{E}-14$ (3.494E-03) | (-7.494E-03, 7.494E-03) | 50.00 | $\begin{gathered} -2.745 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| 2 | Mean | $-3.389 \mathrm{E}-03$ (4.778E-04) | (-4.330E-03, -2.448E-03) | >99.99 | $\begin{gathered} -7.093 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0881 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.741 \mathrm{E}-13$ (3.382E-03) | (-7.255E-03, 7.255E-03) | 50.00 | $\underset{(>0.999)}{5.147 \mathrm{E}-11}$ |  |
| 3 | Mean | $-3.540 \mathrm{E}-03(4.714 \mathrm{E}-04)$ | (-4.468E-03, -2.611E-03) | >99.99 | $\underset{(<0.001)}{-7.509 \mathrm{E}+00}$ | $\begin{gathered} 0.0878 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.174 \mathrm{E}-14$ (3.533E-03) | $(-7.577 \mathrm{E}-03,7.577 \mathrm{E}-03)$ | 50.00 | $\underset{(>0.999)}{-1.182 \mathrm{E}-11}$ |  |
| 4 | Mean | $-3.423 \mathrm{E}-03$ (4.736E-04) | (-4.356E-03, -2.490E-03) | >99.99 | $\begin{gathered} -7.227 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0882 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.590 \mathrm{E}-13$ (3.416E-03) | (-7.327E-03, 7.327E-03) | 50.00 | $\begin{gathered} 4.654 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| 5 | Mean | $-3.621 \mathrm{E}-03(4.681 \mathrm{E}-04)$ | $(-4.543 \mathrm{E}-03,-2.699 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -7.735 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0877 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-5.338 \mathrm{E}-13$ (3.614E-03) | (-7.751E-03, 7.751E-03) | 50.00 | $\begin{gathered} -1.477 \mathrm{E}-10 \\ (>0.999) \end{gathered}$ |  |
| 6 | Mean | $-3.513 \mathrm{E}-03(4.699 \mathrm{E}-04)$ | (-4.438E-03, -2.587E-03) | >99.99 | $\underset{(<0.001)}{-7.475 \mathrm{E}+00}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.688 \mathrm{E}-14$ (3.506E-03) | $(-7.519 \mathrm{E}-03,7.519 \mathrm{E}-03)$ | 50.00 | $\begin{gathered} 4.814 \mathrm{E}-12 \\ (>0.999) \end{gathered}$ |  |
| 7 | Mean | $-3.545 \mathrm{E}-03$ (4.686E-04) | (-4.468E-03, -2.622E-03) | >99.99 | $\begin{gathered} -7.565 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-2.043 \mathrm{E}-13$ (3.538E-03) | (-7.588E-03, 7.588E-03) | 50.00 | $\underset{(>0.999)}{-5.774 \mathrm{E}-11}$ |  |
| 8 | Mean | $-3.457 \mathrm{E}-03$ (4.712E-04) | (-4.385E-03, -2.529E-03) | >99.99 | $\begin{gathered} -7.337 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0881 \\ (<0.0001) \end{gathered}$ |
|  | Median | $1.741 \mathrm{E}-13$ (3.451E-03) | (-7.401E-03, 7.401E-03) | 50.00 | $\begin{gathered} 5.045 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |
| 9 | Mean | $-3.543 \mathrm{E}-03$ (4.688E-04) | (-4.466E-03, -2.620E-03) | >99.99 | $\begin{gathered} -7.557 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0880 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-3.189 \mathrm{E}-13$ (3.536E-03) | $(-7.584 \mathrm{E}-03,7.584 \mathrm{E}-03)$ | 50.00 | $\underset{(>0.999)}{-9.018 \mathrm{E}-11}$ |  |
| Overall | Mean | $-3.504 \mathrm{E}-03$ (4.722E-04) | (-4.433E-03, -2.574E-03) | >99.99 | $\begin{gathered} -7.421 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.0879 \\ (<0.0001) \end{gathered}$ |
|  | Median | $-4.174 \mathrm{E}-14$ (3.506E-03) | (-7.519E-03, 7.519E-03) | 50.00 | $\begin{gathered} -1.182 \mathrm{E}-11 \\ (>0.999) \end{gathered}$ |  |

Table K-13. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\begin{gathered} \text { Tstat } \\ (p-\text { value }) \end{gathered}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-1.334 \mathrm{E}-02(2.593 \mathrm{E}-03)$ | (-1.862E-02, -8.056E-03) | 99.98 | $\begin{gathered} -5.764 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5102 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.363 \mathrm{E}-02(3.771 \mathrm{E}-03)$ | (5.952E-03, 2.131E-02) | 99.90 | $\begin{gathered} 3.820 \mathrm{E}+00 \\ (0.002) \end{gathered}$ | $\begin{gathered} 0.4851 \\ (<0.0001) \end{gathered}$ |
| 1 | Lower | $-1.231 \mathrm{E}-02(1.420 \mathrm{E}-03)$ | $(-1.520 \mathrm{E}-02,-9.417 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -8.669 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5202 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.423 \mathrm{E}-02(4.443 \mathrm{E}-03)$ | (5.176E-03, 2.328E-02) | 99.85 | $\begin{gathered} 3.202 \mathrm{E}+00 \\ (0.003) \end{gathered}$ | $\begin{gathered} 0.4749 \\ (<0.0001) \end{gathered}$ |
| 2 | Lower | $-1.502 \mathrm{E}-02(4.584 \mathrm{E}-03)$ | $(-2.436 \mathrm{E}-02,-5.684 \mathrm{E}-03)$ | 99.87 | $\begin{gathered} -3.277 \mathrm{E}+00 \\ (0.003) \end{gathered}$ | $\begin{gathered} 0.4902 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.224 \mathrm{E}-02(2.132 \mathrm{E}-03)$ | (7.901E-03, 1.659E-02) | >99.99 | $\begin{gathered} 5.743 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5068 \\ (<0.0001) \end{gathered}$ |
| 3 | Lower | $-1.201 \mathrm{E}-02(1.522 \mathrm{E}-03)$ | $(-1.511 \mathrm{E}-02,-8.912 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -7.891 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5220 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.477 \mathrm{E}-02(4.932 \mathrm{E}-03)$ | (4.721E-03, 2.481E-02) | 99.74 | $\begin{gathered} 2.994 \mathrm{E}+00 \\ (0.005) \end{gathered}$ | $\begin{gathered} 0.4722 \\ (<0.0001) \end{gathered}$ |
| 4 | Lower | $-1.415 \mathrm{E}-02(3.795 \mathrm{E}-03)$ | $(-2.188 \mathrm{E}-02,-6.417 \mathrm{E}-03)$ | 99.96 | $\begin{gathered} -3.728 \mathrm{E}+00 \\ (0.001) \end{gathered}$ | $\begin{gathered} 0.5014 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.291 \mathrm{E}-02(2.549 \mathrm{E}-03)$ | (7.717E-03, 1.810E-02) | >99.99 | $\begin{gathered} 5.064 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.4939 \\ (<0.0001) \end{gathered}$ |
| 5 | Lower | $-1.253 \mathrm{E}-02(1.752 \mathrm{E}-03)$ | $(-1.609 \mathrm{E}-02,-8.957 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -7.151 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5197 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.441 \mathrm{E}-02(4.707 \mathrm{E}-03)$ | (4.822E-03, 2.400E-02) | 99.78 | $\begin{gathered} 3.061 \mathrm{E}+00 \\ (0.004) \end{gathered}$ | $\begin{gathered} 0.4751 \\ (<0.0001) \end{gathered}$ |
| 6 | Lower | $-1.330 \mathrm{E}-02(2.606 \mathrm{E}-03)$ | $(-1.860 \mathrm{E}-02,-7.987 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -5.102 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5113 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.362 \mathrm{E}-02(3.574 \mathrm{E}-03)$ | (6.342E-03, 2.090E-02) | 99.97 | $\begin{gathered} 3.811 \mathrm{E}+00 \\ (0.001) \end{gathered}$ | $\begin{gathered} 0.4825 \\ (<0.0001) \end{gathered}$ |
| 7 | Lower | $-1.372 \mathrm{E}-02(2.605 \mathrm{E}-03)$ | $(-1.902 \mathrm{E}-02,-8.412 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -5.266 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5070 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.336 \mathrm{E}-02(3.774 \mathrm{E}-03)$ | (5.669E-03, 2.104E-02) | 99.94 | $\begin{gathered} 3.539 \mathrm{E}+00 \\ (0.001) \end{gathered}$ | $\begin{gathered} 0.4895 \\ (<0.0001) \end{gathered}$ |
| 8 | Lower | $-1.338 \mathrm{E}-02(2.264 \mathrm{E}-03)$ | $(-1.799 \mathrm{E}-02,-8.764 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -5.908 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5113 \\ (<0.0001) \\ \hline \end{gathered}$ |
|  | Upper | $1.371 \mathrm{E}-02$ (4.159E-03) | (5.241E-03, 2.219E-02) | 99.88 | $\begin{gathered} 3.297 \mathrm{E}+00 \\ (0.002) \end{gathered}$ | $\begin{gathered} 0.4834 \\ (<0.0001) \end{gathered}$ |
| 9 | Lower | $-1.364 \mathrm{E}-02(2.792 \mathrm{E}-03)$ | $(-1.933 \mathrm{E}-02,-7.955 \mathrm{E}-03)$ | >99.99 | $\begin{gathered} -4.886 \mathrm{E}+00 \\ (<0.001) \end{gathered}$ | $\begin{gathered} 0.5083 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.345 \mathrm{E}-02(3.668 \mathrm{E}-03)$ | (5.977E-03, 2.092E-02) | 99.96 | $\begin{gathered} 3.666 \mathrm{E}+00 \\ (0.001) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4873 \\ (<0.0001) \end{gathered}$ |

Table K-14. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.168 \mathrm{E}-12(4.511 \mathrm{E}-12)$ | $(-2.169 \mathrm{E}-11,3.356 \mathrm{E}-12)$ | 94.41 | $-2.033 \mathrm{E}+00(0.112)$ |
|  | Upper | $9.329 \mathrm{E}-12$ (4.640E-12) | $(-3.554 \mathrm{E}-12,2.221 \mathrm{E}-11)$ | 94.26 | $2.011 \mathrm{E}+00$ (0.115) |
| 1 | Lower | $-8.706 \mathrm{E}-12(4.254 \mathrm{E}-12)$ | $(-2.052 \mathrm{E}-11,3.105 \mathrm{E}-12)$ | 94.49 | $-2.046 \mathrm{E}+00$ (0.110) |
|  | Upper | $9.584 \mathrm{E}-12$ (4.777E-12) | $(-3.679 \mathrm{E}-12,2.285 \mathrm{E}-11)$ | 94.24 | $2.006 \mathrm{E}+00$ (0.115) |
| 2 | Lower | $-1.000 \mathrm{E}-11(4.956 \mathrm{E}-12)$ | (-2.376E-11, 3.755E-12) | 94.32 | $-2.019 \mathrm{E}+00$ (0.114) |
|  | Upper | $8.609 \mathrm{E}-12$ (4.263E-12) | $(-3.228 \mathrm{E}-12,2.045 \mathrm{E}-11)$ | 94.32 | $2.019 \mathrm{E}+00$ (0.114) |
| 3 | Lower | $-8.748 \mathrm{E}-12$ (4.275E-12) | $(-2.062 \mathrm{E}-11,3.123 \mathrm{E}-12)$ | 94.49 | $-2.046 \mathrm{E}+00$ (0.110) |
|  | Upper | $9.645 \mathrm{E}-12$ (4.812E-12) | $(-3.715 \mathrm{E}-12,2.300 \mathrm{E}-11)$ | 94.22 | $2.004 \mathrm{E}+00$ (0.116) |
| 4 | Lower | $-9.805 \mathrm{E}-12(4.843 \mathrm{E}-12)$ | $(-2.325 \mathrm{E}-11,3.642 \mathrm{E}-12)$ | 94.35 | $-2.025 \mathrm{E}+00(0.113)$ |
|  | Upper | $8.722 \mathrm{E}-12$ (4.326E-12) | $(-3.289 \mathrm{E}-12,2.073 \mathrm{E}-11)$ | 94.30 | $2.016 \mathrm{E}+00(0.114)$ |
| 5 | Lower | $-8.814 \mathrm{E}-12(4.316 \mathrm{E}-12)$ | $(-2.080 \mathrm{E}-11,3.170 \mathrm{E}-12)$ | 94.47 | $-2.042 \mathrm{E}+00(0.111)$ |
|  | Upper | $9.618 \mathrm{E}-12$ (4.795E-12) | $(-3.696 \mathrm{E}-12,2.293 \mathrm{E}-11)$ | 94.23 | $2.006 \mathrm{E}+00(0.115)$ |
| 6 | Lower | $-9.298 \mathrm{E}-12(4.573 \mathrm{E}-12)$ | (-2.199E-11, 3.398E-12) | 94.41 | $-2.033 \mathrm{E}+00(0.112)$ |
|  | Upper | $9.165 \mathrm{E}-12$ (4.558E-12) | $(-3.489 \mathrm{E}-12,2.182 \mathrm{E}-11)$ | 94.27 | $2.011 \mathrm{E}+00(0.115)$ |
| 7 | Lower | $-9.168 \mathrm{E}-12(4.511 \mathrm{E}-12)$ | (-2.169E-11, 3.356E-12) | 94.41 | $-2.033 \mathrm{E}+00(0.112)$ |
|  | Upper | $9.329 \mathrm{E}-12$ (4.640E-12) | (-3.554E-12, 2.221E-11) | 94.26 | $2.011 \mathrm{E}+00(0.115)$ |
| 8 | Lower | $-9.010 \mathrm{E}-12(4.427 \mathrm{E}-12)$ | $(-2.130 \mathrm{E}-11,3.281 \mathrm{E}-12)$ | 94.42 | $-2.035 \mathrm{E}+00(0.112)$ |
|  | Upper | $9.461 \mathrm{E}-12$ (4.710E-12) | $(-3.615 \mathrm{E}-12,2.254 \mathrm{E}-11)$ | 94.25 | $2.009 \mathrm{E}+00(0.115)$ |
| 9 | Lower | $-9.295 \mathrm{E}-12(4.576 \mathrm{E}-12)$ | $(-2.200 \mathrm{E}-11,3.410 \mathrm{E}-12)$ | 94.40 | $-2.031 \mathrm{E}+00(0.112)$ |
|  | Upper | $9.238 \mathrm{E}-12$ (4.594E-12) | $(-3.517 \mathrm{E}-12,2.199 \mathrm{E}-11)$ | 94.27 | $2.011 \mathrm{E}+00(0.115)$ |

Table K-15. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | $\underset{(p-\text { value })}{\text { Tstat }}$ | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | -1.335E-02 (6.017E-03) | (-2.561E-02, -1.095E-03) | 92.66 | $-5.973 \mathrm{E}+00$ (0.147) | $\begin{gathered} 0.5053 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.365 \mathrm{E}-02$ (6.680E-03) | (4.530E-05, 2.726E-02) | 91.74 | $3.849 \mathrm{E}+00$ (0.165) | $\begin{gathered} 0.4818 \\ (<0.0001) \end{gathered}$ |
| 1 | Lower | $-1.214 \mathrm{E}-02(4.750 \mathrm{E}-03)$ | $(-2.181 \mathrm{E}-02,-2.463 \mathrm{E}-03)$ | 93.03 | $-9.316 \mathrm{E}+00$ (0.139) | $\begin{gathered} 0.5129 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.405 \mathrm{E}-02$ (7.010E-03) | $(-2.290 \mathrm{E}-04,2.833 \mathrm{E}-02)$ | 90.98 | $3.324 \mathrm{E}+00$ (0.180) | $\begin{gathered} 0.4731 \\ (<0.0001) \end{gathered}$ |
| 2 | Lower | $-1.500 \mathrm{E}-02(8.071 \mathrm{E}-03)$ | (-3.144E-02, 1.440E-03) | 91.33 | $-3.353 \mathrm{E}+00(0.173)$ | $\begin{gathered} 0.4879 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.216 \mathrm{E}-02$ (5.590E-03) | (7.752E-04, 2.355E-02) | 92.55 | $5.437 \mathrm{E}+00$ (0.149) | $\begin{gathered} 0.4970 \\ (<0.0001) \end{gathered}$ |
| 3 | Lower | $-1.222 \mathrm{E}-02(4.835 \mathrm{E}-03)$ | (-2.207E-02, -2.372E-03) | 93.48 | $-8.930 \mathrm{E}+00$ (0.130) | $\begin{gathered} 0.5187 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.491 \mathrm{E}-02$ (7.421E-03) | (-2.098E-04, 3.002E-02) | 91.21 | $3.233 \mathrm{E}+00$ (0.176) | $\begin{gathered} 0.4724 \\ (<0.0001) \end{gathered}$ |
| 4 | Lower | $-1.400 \mathrm{E}-02(7.135 \mathrm{E}-03)$ | $(-2.854 \mathrm{E}-02,5.307 \mathrm{E}-04)$ | 91.82 | $-3.710 \mathrm{E}+00$ (0.164) | $\begin{gathered} 0.4961 \\ (<0.0001) \\ \hline \end{gathered}$ |
|  | Upper | $1.282 \mathrm{E}-02(5.687 \mathrm{E}-03)$ | (1.234E-03, 2.440E-02) | 92.70 | $4.956 \mathrm{E}+00$ (0.146) | $\begin{gathered} 0.4867 \\ (<0.0001) \end{gathered}$ |
| 5 | Lower | $-1.252 \mathrm{E}-02$ (5.277E-03) | (-2.327E-02, -1.770E-03) | 93.14 | $-7.404 \mathrm{E}+00$ (0.137) | $\begin{gathered} 0.5131 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.457 \mathrm{E}-02$ (7.484E-03) | (-6.692E-04, 2.982E-02) | 91.24 | $3.225 \mathrm{E}+00$ (0.175) | $\begin{gathered} 0.4739 \\ (<0.0001) \end{gathered}$ |
| 6 | Lower | $-1.345 \mathrm{E}-02(5.959 \mathrm{E}-03)$ | (-2.558E-02, -1.307E-03) | 92.77 | $-5.108 \mathrm{E}+00$ (0.145) | $\begin{gathered} 0.5066 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.375 \mathrm{E}-02$ (6.437E-03) | (6.383E-04, 2.686E-02) | 92.08 | $3.914 \mathrm{E}+00$ (0.158) | $\begin{gathered} 0.4809 \\ (<0.0001) \end{gathered}$ |
| 7 | Lower | $-1.368 \mathrm{E}-02(6.071 \mathrm{E}-03)$ | (-2.604E-02, -1.311E-03) | 92.73 | $-5.214 \mathrm{E}+00(0.145)$ | $\begin{gathered} 0.5015 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.333 \mathrm{E}-02$ (6.739E-03) | $(-3.931 \mathrm{E}-04,2.706 \mathrm{E}-02)$ | 91.71 | $3.518 \mathrm{E}+00$ (0.166) | $\begin{gathered} 0.4853 \\ (<0.0001) \end{gathered}$ |
| 8 | Lower | $-1.352 \mathrm{E}-02(5.763 \mathrm{E}-03)$ | (-2.525E-02, -1.779E-03) | 93.12 | $-5.895 \mathrm{E}+00(0.138)$ | $\begin{gathered} 0.5054 \\ (<0.0001) \\ \hline \end{gathered}$ |
|  | Upper | $1.381 \mathrm{E}-02(7.059 \mathrm{E}-03)$ | (-5.706E-04, 2.819E-02) | 91.32 | $3.357 \mathrm{E}+00$ (0.174) | $\begin{gathered} 0.4811 \\ (<0.0001) \end{gathered}$ |
| 9 | Lower | $-1.364 \mathrm{E}-02(6.291 \mathrm{E}-03)$ | (-2.645E-02, -8.264E-04) | 92.54 | $-4.827 \mathrm{E}+00$ (0.149) | $\begin{gathered} 0.5052 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.346 \mathrm{E}-02$ (6.691E-03) | (-1.677E-04, 2.709E-02) | 91.88 | $3.680 \mathrm{E}+00$ (0.162) | $\begin{gathered} 0.4859 \\ (<0.0001) \end{gathered}$ |

Table K-16. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.168 \mathrm{E}-12$ (4.511E-12) | $(-2.169 \mathrm{E}-11,3.285 \mathrm{E}-12)$ | 94.49 | $-2.046 \mathrm{E}+00(0.110)$ |
|  | Upper | $9.329 \mathrm{E}-12$ (4.640E-12) | $(-3.554 \mathrm{E}-12,3.125 \mathrm{E}-11)$ | 94.35 | $2.023 \mathrm{E}+00$ (0.113) |
| 1 | Lower | -8.706E-12 (4.254E-12) | $(-2.052 \mathrm{E}-11,3.031 \mathrm{E}-12)$ | 95.19 | $-2.166 \mathrm{E}+00(0.096)$ |
|  | Upper | $9.584 \mathrm{E}-12(4.777 \mathrm{E}-12)$ | $(-3.679 \mathrm{E}-12,3.207 \mathrm{E}-11)$ | 94.40 | $2.032 \mathrm{E}+00$ (0.112) |
| 2 | Lower | $-1.000 \mathrm{E}-11(4.956 \mathrm{E}-12)$ | $(-2.376 \mathrm{E}-11,3.659 \mathrm{E}-12)$ | 94.41 | $-2.033 \mathrm{E}+00(0.112)$ |
|  | Upper | $8.609 \mathrm{E}-12$ (4.263E-12) | $(-3.228 \mathrm{E}-12,2.871 \mathrm{E}-11)$ | 94.34 | $2.023 \mathrm{E}+00$ (0.113) |
| 3 | Lower | $-8.748 \mathrm{E}-12(4.275 \mathrm{E}-12)$ | $(-2.062 \mathrm{E}-11,3.073 \mathrm{E}-12)$ | 94.56 | $-2.057 \mathrm{E}+00(0.109)$ |
|  | Upper | $9.645 \mathrm{E}-12$ (4.812E-12) | $(-3.715 \mathrm{E}-12,3.231 \mathrm{E}-11)$ | 94.37 | $2.027 \mathrm{E}+00$ (0.113) |
| 4 | Lower | $-9.805 \mathrm{E}-12(4.843 \mathrm{E}-12)$ | $(-2.325 \mathrm{E}-11,3.556 \mathrm{E}-12)$ | 94.44 | $-2.038 \mathrm{E}+00(0.111)$ |
|  | Upper | $8.722 \mathrm{E}-12$ (4.326E-12) | $(-3.289 \mathrm{E}-12,2.079 \mathrm{E}-11)$ | 94.33 | $2.021 \mathrm{E}+00(0.113)$ |
| 5 | Lower | $-8.814 \mathrm{E}-12(4.316 \mathrm{E}-12)$ | $(-2.080 \mathrm{E}-11,3.102 \mathrm{E}-12)$ | 95.23 | $-2.174 \mathrm{E}+00(0.095)$ |
|  | Upper | $9.618 \mathrm{E}-12$ (4.795E-12) | $(-3.696 \mathrm{E}-12,3.221 \mathrm{E}-11)$ | 94.38 | $2.029 \mathrm{E}+00$ (0.112) |
| 6 | Lower | $-9.298 \mathrm{E}-12(4.573 \mathrm{E}-12)$ | $(-2.199 \mathrm{E}-11,3.326 \mathrm{E}-12)$ | 94.49 | $-2.046 \mathrm{E}+00(0.110)$ |
|  | Upper | $9.165 \mathrm{E}-12$ (4.558E-12) | $(-3.482 \mathrm{E}-12,3.009 \mathrm{E}-11)$ | 94.35 | $2.023 \mathrm{E}+00(0.113)$ |
| 7 | Lower | $-9.168 \mathrm{E}-12(4.511 \mathrm{E}-12)$ | $(-2.169 \mathrm{E}-11,3.285 \mathrm{E}-12)$ | 94.49 | $-2.046 \mathrm{E}+00(0.110)$ |
|  | Upper | $9.329 \mathrm{E}-12$ (4.640E-12) | $(-3.554 \mathrm{E}-12,3.125 \mathrm{E}-11)$ | 94.34 | $2.022 \mathrm{E}+00(0.113)$ |
| 8 | Lower | $-9.010 \mathrm{E}-12(4.427 \mathrm{E}-12)$ | $(-2.130 \mathrm{E}-11,3.217 \mathrm{E}-12)$ | 94.50 | $-2.048 \mathrm{E}+00(0.110)$ |
|  | Upper | $9.461 \mathrm{E}-12(4.710 \mathrm{E}-12)$ | $(-3.615 \mathrm{E}-12,3.170 \mathrm{E}-11)$ | 94.34 | $2.023 \mathrm{E}+00(0.113)$ |
| 9 | Lower | $-9.295 \mathrm{E}-12(4.576 \mathrm{E}-12)$ | $(-2.200 \mathrm{E}-11,3.340 \mathrm{E}-12)$ | 94.48 | $-2.044 \mathrm{E}+00(0.110)$ |
|  | Upper | $9.238 \mathrm{E}-12$ (4.594E-12) | $(-3.504 \mathrm{E}-12,2.206 \mathrm{E}-11)$ | 94.39 | $2.030 \mathrm{E}+00$ (0.112) |

Table K-17. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-1.334 \mathrm{E}-02(1.668 \mathrm{E}-03)$ | (-1.674E-02, -9.941E-03) | >99.99 | $-7.989 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5100 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.363 \mathrm{E}-02(1.670 \mathrm{E}-03)$ | (1.023E-02, 1.703E-02) | >99.99 | $8.205 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4848 \\ (<0.0001) \end{gathered}$ |
| 1 | Lower | $-1.231 \mathrm{E}-02(1.567 \mathrm{E}-03)$ | $(-1.550 \mathrm{E}-02,-9.117 \mathrm{E}-03)$ | >99.99 | $-7.854 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5202 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.423 \mathrm{E}-02(1.572 \mathrm{E}-03)$ | (1.102E-02, 1.743E-02) | >99.99 | $9.051 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4747 \\ (<0.0001) \end{gathered}$ |
| 2 | Lower | $-1.502 \mathrm{E}-02(1.837 \mathrm{E}-03)$ | $(-1.876 \mathrm{E}-02,-1.128 \mathrm{E}-02)$ | >99.99 | $-8.175 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4900 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.224 \mathrm{E}-02(1.833 \mathrm{E}-03)$ | (8.509E-03, 1.598E-02) | >99.99 | $6.679 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5063 \\ (<0.0001) \end{gathered}$ |
| 3 | Lower | $-1.201 \mathrm{E}-02(1.546 \mathrm{E}-03)$ | $(-1.516 \mathrm{E}-02,-8.863 \mathrm{E}-03)$ | >99.99 | $-7.768 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5222 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.477 \mathrm{E}-02(1.552 \mathrm{E}-03)$ | (1.161E-02, 1.793E-02) | >99.99 | $9.515 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4720 \\ (<0.0001) \end{gathered}$ |
| 4 | Lower | $-1.415 \mathrm{E}-02(1.747 \mathrm{E}-03)$ | $(-1.771 \mathrm{E}-02,-1.059 \mathrm{E}-02)$ | >99.99 | $-8.099 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5012 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.291 \mathrm{E}-02$ (1.745E-03) | (9.355E-03, 1.646E-02) | >99.99 | $7.399 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4936 \\ (<0.0001) \end{gathered}$ |
| 5 | Lower | $-1.253 \mathrm{E}-02(1.589 \mathrm{E}-03)$ | $(-1.576 \mathrm{E}-02,-9.288 \mathrm{E}-03)$ | >99.99 | $-7.881 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5197 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.441 \mathrm{E}-02(1.594 \mathrm{E}-03)$ | (1.116E-02, 1.766E-02) | >99.99 | $9.039 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4749 \\ (<0.0001) \end{gathered}$ |
| 6 | Lower | $-1.330 \mathrm{E}-02(1.662 \mathrm{E}-03)$ | $(-1.668 \mathrm{E}-02,-9.910 \mathrm{E}-03)$ | >99.99 | $-7.999 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5112 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.362 \mathrm{E}-02(1.664 \mathrm{E}-03)$ | (1.023E-02, 1.701E-02) | >99.99 | $8.189 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4823 \\ (<0.0001) \end{gathered}$ |
| 7 | Lower | $-1.372 \mathrm{E}-02$ (1.701E-03) | $(-1.718 \mathrm{E}-02,-1.025 \mathrm{E}-02)$ | >99.99 | $-8.062 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5068 \\ (<0.0001) \\ \hline \end{gathered}$ |
|  | Upper | $1.336 \mathrm{E}-02(1.703 \mathrm{E}-03)$ | (9.887E-03, 1.683E-02) | >99.99 | $7.841 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4892 \\ (<0.0001) \end{gathered}$ |
| 8 | Lower | $-1.338 \mathrm{E}-02(1.668 \mathrm{E}-03)$ | $(-1.677 \mathrm{E}-02,-9.980 \mathrm{E}-03)$ | >99.99 | $-8.021 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5111 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.371 \mathrm{E}-02(1.671 \mathrm{E}-03)$ | (1.031E-02, 1.712E-02) | >99.99 | $8.208 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4832 \\ (<0.0001) \end{gathered}$ |
| 9 | Lower | $-1.364 \mathrm{E}-02(1.695 \mathrm{E}-03)$ | $(-1.710 \mathrm{E}-02,-1.019 \mathrm{E}-02)$ | >99.99 | $-8.047 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.5081 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.345 \mathrm{E}-02(1.697 \mathrm{E}-03)$ | (9.992E-03, 1.691E-02) | >99.99 | $7.925 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.4871 \\ (<0.0001) \end{gathered}$ |

Table K-18. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Linear Trapezoid Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.164 \mathrm{E}-12(1.317 \mathrm{E}-02)$ | $(-3.657 \mathrm{E}-02,3.657 \mathrm{E}-02)$ | 50.00 | $-7.037 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.389 \mathrm{E}-11$ (1.342E-02) | (-3.725E-02, 3.725E-02) | 50.00 | $1.020 \mathrm{E}-09$ (>0.999) |
| 1 | Lower | $-8.700 \mathrm{E}-12$ (1.212E-02) | $(-3.365 \mathrm{E}-02,3.365 \mathrm{E}-02)$ | 50.00 | $-7.178 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.428 \mathrm{E}-11$ (1.401E-02) | $(-3.889 \mathrm{E}-02,3.889 \mathrm{E}-02)$ | 50.00 | $1.019 \mathrm{E}-09(>0.999)$ |
| 2 | Lower | $-1.000 \mathrm{E}-11(1.479 \mathrm{E}-02)$ | (-4.107E-02, 4.107E-02) | 50.00 | $-6.763 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.279 \mathrm{E}-11$ (1.206E-02) | $(-3.347 \mathrm{E}-02,3.347 \mathrm{E}-02)$ | 50.00 | $1.061 \mathrm{E}-09$ (>0.999) |
| 3 | Lower | $-8.742 \mathrm{E}-12$ (1.183E-02) | (-3.284E-02, 3.284E-02) | 50.00 | $-7.390 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.438 \mathrm{E}-11(1.454 \mathrm{E}-02)$ | (-4.037E-02, 4.037E-02) | 50.00 | $9.891 \mathrm{E}-10$ (>0.999) |
| 4 | Lower | $-9.803 \mathrm{E}-12(1.393 \mathrm{E}-02)$ | $(-3.868 \mathrm{E}-02,3.868 \mathrm{E}-02)$ | $50.00$ | -7.037E-10 (>0.999) |
|  | Upper | $1.296 \mathrm{E}-11(1.271 \mathrm{E}-02)$ | (-3.529E-02, 3.529E-02) | 50.00 | $1.020 \mathrm{E}-09(>0.999)$ |
| 5 | Lower | $-8.808 \mathrm{E}-12$ (1.233E-02) | $(-3.424 \mathrm{E}-02,3.424 \mathrm{E}-02)$ | 50.00 | $-7.142 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.434 \mathrm{E}-11(1.419 \mathrm{E}-02)$ | $(-3.940 \mathrm{E}-02,3.940 \mathrm{E}-02)$ | 50.00 | $1.010 \mathrm{E}-09$ (>0.999) |
| 6 | Lower | $-9.294 \mathrm{E}-12(1.309 \mathrm{E}-02)$ | $(-3.635 \mathrm{E}-02,3.635 \mathrm{E}-02)$ | 50.00 | $-7.099 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.364 \mathrm{E}-11(1.342 \mathrm{E}-02)$ | $(-3.725 \mathrm{E}-02,3.725 \mathrm{E}-02)$ | 50.00 | $1.017 \mathrm{E}-09$ (>0.999) |
| 7 | Lower | $-9.164 \mathrm{E}-12(1.351 \mathrm{E}-02)$ | $(-3.751 \mathrm{E}-02,3.751 \mathrm{E}-02)$ | 50.00 | $-6.784 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.389 \mathrm{E}-11(1.315 \mathrm{E}-02)$ | $(-3.652 \mathrm{E}-02,3.652 \mathrm{E}-02)$ | 50.00 | $1.056 \mathrm{E}-09(>0.999)$ |
| 8 | Lower | $-9.006 \mathrm{E}-12(1.317 \mathrm{E}-02)$ | $(-3.657 \mathrm{E}-02,3.657 \mathrm{E}-02)$ | 50.00 | $-6.837 \mathrm{E}-10$ (>0.999) |
|  | Upper | $1.409 \mathrm{E}-11(1.350 \mathrm{E}-02)$ | $(-3.749 \mathrm{E}-02,3.749 \mathrm{E}-02)$ | 50.00 | $1.044 \mathrm{E}-09$ (>0.999) |
| 9 | Lower | $-9.292 \mathrm{E}-12(1.343 \mathrm{E}-02)$ | $(-3.730 \mathrm{E}-02,3.730 \mathrm{E}-02)$ | 50.00 | $-6.916 \mathrm{E}-10(>0.999)$ |
|  | Upper | $1.375 \mathrm{E}-11$ (1.324E-02) | (-3.677E-02, 3.677E-02) | 50.00 | $1.039 \mathrm{E}-09$ (>0.999) |

Table K-19. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Spline Approximation for $n=30$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat ( $\boldsymbol{p}$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $\begin{aligned} & -2.291 \mathrm{E}-02 \\ & (6.582 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.631 \mathrm{E}-02, \\ & -9.499 \mathrm{E}-03) \end{aligned}$ | 99.91 | $-3.487 \mathrm{E}+00$ (0.002) | 0.2815 (<0.0001) |
|  | Upper | $\begin{gathered} 5.887 \mathrm{E}-03 \\ (9.494 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.345 \mathrm{E}-02 \\ 2.523 \mathrm{E}-02) \end{gathered}$ | 73.30 | $6.359 \mathrm{E}-01$ (0.534) | $0.4441(<0.0001)$ |
| 1 | Lower | $\begin{aligned} & -2.187 \mathrm{E}-02 \\ & (7.148 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.643 \mathrm{E}-02, \\ & -7.308 \mathrm{E}-03) \\ & \hline \end{aligned}$ | 99.78 | $-3.059 \mathrm{E}+00(0.004)$ | $0.2928(<0.0001)$ |
|  | Upper | $\begin{gathered} 6.473 \mathrm{E}-03 \\ (6.974 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-7.732 \mathrm{E}-03 \\ 2.068 \mathrm{E}-02) \end{gathered}$ | 81.99 | $9.282 \mathrm{E}-01$ (0.360) | $0.4363(<0.0001)$ |
| 2 | Lower | $\begin{aligned} & -2.466 \mathrm{E}-02 \\ & (6.898 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.871 \mathrm{E}-02, \\ & -1.061 \mathrm{E}-02) \\ & \hline \end{aligned}$ | 99.94 | $-3.575 \mathrm{E}+00(0.001)$ | $0.2639(<0.0001)$ |
|  | Upper | $\begin{gathered} 4.440 \mathrm{E}-03 \\ (1.093 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.783 \mathrm{E}-02, \\ 2.671 \mathrm{E}-02) \end{gathered}$ | 65.63 | $4.061 \mathrm{E}-01$ (0.687) | $0.4609(<0.0001)$ |
| 3 | Lower | $\begin{aligned} & -2.153 \mathrm{E}-02 \\ & (6.588 \mathrm{E}-03) \\ & \hline \end{aligned}$ | $\begin{aligned} & (-3.495 \mathrm{E}-02, \\ & -8.108 \mathrm{E}-03) \\ & \hline \end{aligned}$ | 99.87 | $-3.268 \mathrm{E}+00(0.003)$ | 0.2980 (<0.0001) |
|  | Upper | $\begin{gathered} 7.068 \mathrm{E}-03 \\ (8.800 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.086 \mathrm{E}-02, \\ 2.499 \mathrm{E}-02) \end{gathered}$ | 78.61 | $8.032 \mathrm{E}-01$ (0.428) | $0.4302(<0.0001)$ |
| 4 | Lower | $\begin{aligned} & -2.376 \mathrm{E}-02 \\ & (6.203 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.640 \mathrm{E}-02, \\ & -1.113 \mathrm{E}-02) \\ & \hline \end{aligned}$ | 99.97 | $-3.831 \mathrm{E}+00(0.001)$ | $0.2733(<0.0001)$ |
|  | Upper | $\begin{gathered} 5.118 \mathrm{E}-03 \\ (9.555 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.435 \mathrm{E}-02, \\ 2.458 \mathrm{E}-02) \end{gathered}$ | 70.20 | $5.356 \mathrm{E}-01$ (0.596) | $0.4517(<0.0001)$ |
| 5 | Lower | $\begin{aligned} & -2.205 \mathrm{E}-02 \\ & (6.514 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.532 \mathrm{E}-02, \\ & -8.777 \mathrm{E}-03) \end{aligned}$ | 99.90 | $-3.384 \mathrm{E}+00(0.002)$ | $0.2905(<0.0001)$ |
|  | Upper | $\begin{gathered} 6.709 \mathrm{E}-03 \\ (9.545 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.273 \mathrm{E}-02 \\ 2.615 \mathrm{E}-02) \end{gathered}$ | 75.64 | $7.029 \mathrm{E}-01$ (0.487) | $0.4349(<0.0001)$ |
| 6 | Lower | $\begin{aligned} & -2.287 \mathrm{E}-02 \\ & (6.214 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.553 \mathrm{E}-02, \\ & -1.021 \mathrm{E}-02) \\ & \hline \end{aligned}$ | 99.96 | $-3.681 \mathrm{E}+00(0.001)$ | $0.2813(<0.0001)$ |
|  | Upper | $\begin{gathered} 5.868 \mathrm{E}-03 \\ (9.059 \mathrm{E}-03) \end{gathered}$ | $\begin{gathered} (-1.258 \mathrm{E}-02, \\ 2.432 \mathrm{E}-02) \end{gathered}$ | 73.91 | $6.478 \mathrm{E}-01$ (0.522) | $0.4437(<0.0001)$ |
| 7 | Lower | $\begin{aligned} & -2.329 \mathrm{E}-02 \\ & (6.449 \mathrm{E}-03) \\ & \hline \end{aligned}$ | $\begin{aligned} & (-3.642 \mathrm{E}-02, \\ & -1.015 \mathrm{E}-02) \\ & \hline \end{aligned}$ | 99.95 | $-3.611 \mathrm{E}+00(0.001)$ | $0.2765(<0.0001)$ |
|  | Upper | $\begin{gathered} 5.612 \mathrm{E}-03 \\ (1.005 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.485 \mathrm{E}-02, \\ 2.608 \mathrm{E}-02) \end{gathered}$ | 70.98 | $5.585 \mathrm{E}-01$ (0.580) | $0.4484(<0.0001)$ |
| 8 | Lower | $\begin{aligned} & -2.293 \mathrm{E}-02 \\ & (6.527 \mathrm{E}-03) \\ & \hline \end{aligned}$ | $\begin{aligned} & (-3.622 \mathrm{E}-02, \\ & -9.632 \mathrm{E}-03) \\ & \hline \end{aligned}$ | 99.93 | $-3.513 \mathrm{E}+00(0.001)$ | $0.2798(<0.0001)$ |
|  | Upper | $\begin{gathered} 5.990 \mathrm{E}-03 \\ (1.009 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.455 \mathrm{E}-02 \\ 2.653 \mathrm{E}-02) \end{gathered}$ | 72.16 | $5.939 \mathrm{E}-01$ (0.557) | $0.4439(<0.0001)$ |
| 9 | Lower | $\begin{aligned} & -2.321 \mathrm{E}-02 \\ & (6.697 \mathrm{E}-03) \end{aligned}$ | $\begin{aligned} & (-3.685 \mathrm{E}-02, \\ & -9.569 \mathrm{E}-03) \\ & \hline \end{aligned}$ | 99.92 | $-3.466 \mathrm{E}+00(0.002)$ | $0.2778(<0.0001)$ |
|  | Upper | $\begin{gathered} 5.709 \mathrm{E}-03 \\ (1.044 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.556 \mathrm{E}-02, \\ 2.698 \mathrm{E}-02) \end{gathered}$ | 70.58 | $5.466 \mathrm{E}-01$ (0.588) | $0.4469(<0.0001)$ |

Table K-20. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Spline Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.162 \mathrm{E}-12(4.818 \mathrm{E}-14)$ | $(-9.293 \mathrm{E}-12,-9.030 \mathrm{E}-12)$ | >99.99 | $-1.906 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.294 \mathrm{E}-12$ (4.597E-12) | $(-3.469 \mathrm{E}-12,2.206 \mathrm{E}-11)$ | 94.34 | $2.023 \mathrm{E}+00$ (0.113) |
| 1 | Lower | $-8.695 \mathrm{E}-12(5.667 \mathrm{E}-14)$ | $(-8.852 \mathrm{E}-12,-8.538 \mathrm{E}-12)$ | >99.99 | $-1.534 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.563 \mathrm{E}-12$ (4.718E-12) | $(-3.535 \mathrm{E}-12,2.266 \mathrm{E}-11)$ | 94.37 | $2.027 \mathrm{E}+00$ (0.113) |
| 2 | Lower | $-1.000 \mathrm{E}-11$ (3.966E-14) | $(-1.011 \mathrm{E}-11,-9.893 \mathrm{E}-12)$ | >99.99 | $-2.522 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.550 \mathrm{E}-12$ (4.238E-12) | $(-3.215 \mathrm{E}-12,2.032 \mathrm{E}-11)$ | 94.31 | $2.018 \mathrm{E}+00(0.114)$ |
| 3 | Lower | $-8.735 \mathrm{E}-12(5.491 \mathrm{E}-14)$ | $(-8.887 \mathrm{E}-12,-8.583 \mathrm{E}-12)$ | >99.99 | $-1.591 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.630 \mathrm{E}-12$ (4.752E-12) | $(-3.564 \mathrm{E}-12,2.282 \mathrm{E}-11)$ | 94.37 | $2.027 \mathrm{E}+00$ (0.113) |
| 4 | Lower | $-9.801 \mathrm{E}-12(4.463 \mathrm{E}-14)$ | $(-9.925 \mathrm{E}-12,-9.677 \mathrm{E}-12)$ | >99.99 | $-2.196 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.673 \mathrm{E}-12$ (4.291E-12) | $(-3.242 \mathrm{E}-12,2.059 \mathrm{E}-11)$ | 94.33 | $2.021 \mathrm{E}+00(0.113)$ |
| 5 | Lower | $-8.803 \mathrm{E}-12(5.209 \mathrm{E}-14)$ | $(-8.948 \mathrm{E}-12,-8.658 \mathrm{E}-12)$ | >99.99 | $-1.690 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.598 \mathrm{E}-12(4.740 \mathrm{E}-12)$ | $(-3.562 \mathrm{E}-12,2.276 \mathrm{E}-11)$ | $94.36$ | $2.025 \mathrm{E}+00(0.113)$ |
| 6 | Lower | $-9.292 \mathrm{E}-12(4.874 \mathrm{E}-14)$ | $(-9.427 \mathrm{E}-12,-9.156 \mathrm{E}-12)$ | >99.99 | $-1.906 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.130 \mathrm{E}-12$ (4.512E-12) | $(-3.398 \mathrm{E}-12,2.166 \mathrm{E}-11)$ | 94.35 | $2.023 \mathrm{E}+00$ (0.113) |
| 7 | Lower | $-9.162 \mathrm{E}-12(4.730 \mathrm{E}-14)$ | $(-9.293 \mathrm{E}-12,-9.030 \mathrm{E}-12)$ | >99.99 | $-1.937 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.294 \mathrm{E}-12$ (4.597E-12) | $(-3.469 \mathrm{E}-12,2.206 \mathrm{E}-11)$ | 94.34 | $2.022 \mathrm{E}+00(0.113)$ |
| 8 | Lower | $-9.002 \mathrm{E}-12(4.818 \mathrm{E}-14)$ | $(-9.136 \mathrm{E}-12,-8.869 \mathrm{E}-12)$ | >99.99 | $-1.868 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.431 \mathrm{E}-12(4.663 \mathrm{E}-12)$ | $(-3.514 \mathrm{E}-12,2.238 \mathrm{E}-11)$ | 94.34 | $2.023 \mathrm{E}+00(0.113)$ |
| 9 | Lower | $-9.289 \mathrm{E}-12(4.632 \mathrm{E}-14)$ | (-9.418E-12, -9.161E-12) | >99.99 | $-2.006 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.202 \mathrm{E}-12$ (4.551E-12) | $(-3.434 \mathrm{E}-12,2.184 \mathrm{E}-11)$ | 94.34 | $2.022 \mathrm{E}+00$ (0.113) |

Table K-21. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Spline Approximation for $n=30$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $\begin{aligned} & -2.290 \mathrm{E}-02 \\ & (1.091 \mathrm{E}-02) \\ & \hline \end{aligned}$ | $\begin{aligned} & (-4.513 \mathrm{E}-02, \\ & -6.668 \mathrm{E}-04) \end{aligned}$ | 86.48 | $\begin{gathered} -2.875 \mathrm{E}+00 \\ (0.270) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4050 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 5.947 \mathrm{E}-03 \\ (1.182 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.814 \mathrm{E}-02, \\ 3.003 \mathrm{E}-02) \\ \hline \end{gathered}$ | 80.08 | $\begin{gathered} 2.214 \mathrm{E}+00 \\ (0.398) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4479 \\ (<0.0001) \end{gathered}$ |
| 1 | Lower | $\begin{aligned} & -2.170 \mathrm{E}-02 \\ & (1.137 \mathrm{E}-02) \\ & \hline \end{aligned}$ | $\begin{gathered} (-4.487 \mathrm{E}-02, \\ 1.458 \mathrm{E}-03) \\ \hline \end{gathered}$ | 85.00 | $\begin{gathered} -4.903 \mathrm{E}+00 \\ (0.300) \end{gathered}$ | $\begin{gathered} 0.4062 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 6.507 \mathrm{E}-03 \\ (1.092 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.574 \mathrm{E}-02, \\ 2.875 \mathrm{E}-02) \\ \hline \end{gathered}$ | 82.95 | $\begin{gathered} 1.033 \mathrm{E}+00 \\ (0.341) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4430 \\ (<0.0001) \end{gathered}$ |
| 2 | Lower | $\begin{aligned} & -2.473 \mathrm{E}-02 \\ & (1.083 \mathrm{E}-02) \end{aligned}$ | $\begin{aligned} & (-4.680 \mathrm{E}-02, \\ & -2.661 \mathrm{E}-03) \end{aligned}$ | 87.10 | $\begin{gathered} -2.545 \mathrm{E}+00 \\ (0.258) \end{gathered}$ | $\begin{gathered} 0.3953 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 4.491 \mathrm{E}-03 \\ (1.253 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-2.104 \mathrm{E}-02, \\ 3.002 \mathrm{E}-02) \\ \hline \end{gathered}$ | 78.02 | $\begin{gathered} 2.565 \mathrm{E}+00 \\ (0.440) \end{gathered}$ | $\begin{gathered} 0.4567 \\ (<0.0001) \end{gathered}$ |
| 3 | Lower | $\begin{aligned} & -2.174 \mathrm{E}-02 \\ & (1.118 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-4.451 \mathrm{E}-02, \\ 1.035 \mathrm{E}-03) \\ \hline \end{gathered}$ | 86.43 | $\begin{gathered} -2.768 \mathrm{E}+00 \\ (0.271) \end{gathered}$ | $\begin{gathered} 0.4078 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 7.055 \mathrm{E}-03 \\ (1.167 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.673 \mathrm{E}-02, \\ 3.084 \mathrm{E}-02) \\ \hline \end{gathered}$ | 81.11 | $\begin{gathered} 1.749 \mathrm{E}+00 \\ (0.378) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4420 \\ (<0.0001) \end{gathered}$ |
| 4 | Lower | $\begin{aligned} & -2.370 \mathrm{E}-02 \\ & (1.066 \mathrm{E}-02) \\ & \hline \end{aligned}$ | $\begin{aligned} & (-4.542 \mathrm{E}-02, \\ & -1.984 \mathrm{E}-03) \end{aligned}$ | 86.49 | $\begin{gathered} -2.762 \mathrm{E}+00 \\ (0.270) \end{gathered}$ | $\begin{gathered} 0.4037 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 5.202 \mathrm{E}-03 \\ (1.172 \mathrm{E}-02) \\ \hline \end{gathered}$ | $\begin{gathered} (-1.866 \mathrm{E}-02, \\ 2.907 \mathrm{E}-02) \\ \hline \end{gathered}$ | 79.66 | $\begin{gathered} 2.034 \mathrm{E}+00 \\ (0.407) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4506 \\ (<0.0001) \end{gathered}$ |
| 5 | Lower | $\begin{aligned} & -2.215 \mathrm{E}-02 \\ & (1.095 \mathrm{E}-02) \end{aligned}$ | $\begin{gathered} (-4.446 \mathrm{E}-02, \\ 1.493 \mathrm{E}-04) \end{gathered}$ | 86.81 | $\begin{gathered} -2.548 \mathrm{E}+00 \\ (0.264) \end{gathered}$ | $\begin{gathered} 0.4072 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 6.797 \mathrm{E}-03 \\ (1.181 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.725 \mathrm{E}-02, \\ 3.085 \mathrm{E}-02) \\ \hline \end{gathered}$ | 80.55 | $\begin{gathered} 2.364 \mathrm{E}+00 \\ (0.389) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4435 \\ (<0.0001) \\ \hline \end{gathered}$ |
| 6 | Lower | $\begin{aligned} & -2.274 \mathrm{E}-02 \\ & (1.081 \mathrm{E}-02) \end{aligned}$ | $\begin{aligned} & (-4.476 \mathrm{E}-02, \\ & -7.244 \mathrm{E}-04) \end{aligned}$ | 86.18 | $\begin{gathered} -2.809 \mathrm{E}+00 \\ (0.276) \end{gathered}$ | $\begin{gathered} 0.4047 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 5.869 \mathrm{E}-03 \\ (1.158 \mathrm{E}-02) \\ \hline \end{gathered}$ | $\begin{gathered} (-1.773 \mathrm{E}-02, \\ 2.947 \mathrm{E}-02) \\ \hline \end{gathered}$ | 80.22 | $\begin{gathered} 1.768 \mathrm{E}+00 \\ (0.396) \end{gathered}$ | $\begin{gathered} 0.4476 \\ (<0.0001) \end{gathered}$ |
| 7 | Lower | $\begin{aligned} & -2.348 \mathrm{E}-02 \\ & (1.066 \mathrm{E}-02) \\ & \hline \end{aligned}$ | $\begin{gathered} (-4.519 \mathrm{E}-02, \\ -1.780 \mathrm{E}-03) \end{gathered}$ | 86.89 | $\begin{gathered} -2.592 \mathrm{E}+00 \\ (0.262) \end{gathered}$ | $\begin{gathered} 0.4048 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 5.833 \mathrm{E}-03 \\ (1.188 \mathrm{E}-02) \\ \hline \end{gathered}$ | $\begin{gathered} (-1.837 \mathrm{E}-02, \\ 3.004 \mathrm{E}-02) \\ \hline \end{gathered}$ | 79.72 | $\begin{gathered} 2.886 \mathrm{E}+00 \\ (0.406) \end{gathered}$ | $\begin{gathered} 0.4497 \\ (<0.0001) \end{gathered}$ |
| 8 | Lower | $\begin{aligned} & -2.246 \mathrm{E}-02 \\ & (1.076 \mathrm{E}-02) \end{aligned}$ | $\begin{aligned} & (-4.437 \mathrm{E}-02, \\ & -5.413 \mathrm{E}-04) \end{aligned}$ | 86.54 | $\begin{gathered} -2.483 \mathrm{E}+00 \\ (0.269) \end{gathered}$ | $\begin{gathered} 0.4091 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 6.164 \mathrm{E}-03 \\ (1.189 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.806 \mathrm{E}-02, \\ 3.038 \mathrm{E}-02) \\ \hline \end{gathered}$ | 79.62 | $\begin{gathered} 2.904 \mathrm{E}+00 \\ (0.408) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4480 \\ (<0.0001) \end{gathered}$ |
| 9 | Lower | $\begin{aligned} & -2.338 \mathrm{E}-02 \\ & (1.101 \mathrm{E}-02) \\ & \hline \end{aligned}$ | $\begin{aligned} & (-4.580 \mathrm{E}-02, \\ & -9.523 \mathrm{E}-04) \end{aligned}$ | 86.86 | $\begin{gathered} -2.465 \mathrm{E}+00 \\ (0.263) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4065 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $\begin{gathered} 5.606 \mathrm{E}-03 \\ (1.240 \mathrm{E}-02) \end{gathered}$ | $\begin{gathered} (-1.965 \mathrm{E}-02, \\ 3.087 \mathrm{E}-02) \\ \hline \end{gathered}$ | 78.82 | $\begin{gathered} 2.625 \mathrm{E}+00 \\ (0.424) \end{gathered}$ | $\begin{gathered} 0.4502 \\ (<0.0001) \end{gathered}$ |

Table K-22. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Spline Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.162 \mathrm{E}-12(3.654 \mathrm{E}-12)$ | $(-1.899 \mathrm{E}-11,3.437 \mathrm{E}-13)$ | 97.24 | $-2.681 \mathrm{E}+00(0.055)$ |
|  | Upper | $9.294 \mathrm{E}-12$ (4.597E-12) | $(-3.435 \mathrm{E}-12,2.213 \mathrm{E}-11)$ | 96.41 | $2.433 \mathrm{E}+00$ (0.072) |
| 1 | Lower | $-8.695 \mathrm{E}-12$ (3.743E-12) | $(-1.899 \mathrm{E}-11,1.704 \mathrm{E}-12)$ | 95.97 | $-2.327 \mathrm{E}+00(0.081)$ |
|  | Upper | $9.563 \mathrm{E}-12$ (4.718E-12) | $(-3.514 \mathrm{E}-12,2.269 \mathrm{E}-11)$ | 96.50 | $2.457 \mathrm{E}+00$ (0.070) |
| 2 | Lower | $-1.000 \mathrm{E}-11(3.116 \mathrm{E}-12)$ | $(-1.858 \mathrm{E}-11,-1.369 \mathrm{E}-12)$ | 98.39 | $-3.221 \mathrm{E}+00(0.032)$ |
|  | Upper | $8.550 \mathrm{E}-12$ (4.238E-12) | $(-3.156 \mathrm{E}-12,2.037 \mathrm{E}-11)$ | 95.28 | $2.184 \mathrm{E}+00$ (0.094) |
| 3 | Lower | $-8.735 \mathrm{E}-12(4.104 \mathrm{E}-12)$ | $(-2.003 \mathrm{E}-11,2.679 \mathrm{E}-12)$ | 94.99 | $-2.130 \mathrm{E}+00$ (0.100) |
|  | Upper | $9.630 \mathrm{E}-12$ (4.752E-12) | $(-3.548 \mathrm{E}-12,2.284 \mathrm{E}-11)$ | 96.51 | $2.458 \mathrm{E}+00$ (0.070) |
| 4 | Lower | $-9.801 \mathrm{E}-12$ (3.654E-12) | (-1.986E-11, 3.437E-13) | 97.24 | $-2.681 \mathrm{E}+00(0.055)$ |
|  | Upper | $8.673 \mathrm{E}-12$ (4.291E-12) | $(-3.193 \mathrm{E}-12,2.064 \mathrm{E}-11)$ | 96.30 | $2.298 \mathrm{E}+00$ (0.074) |
| 5 | Lower | $-8.803 \mathrm{E}-12(3.710 \mathrm{E}-12)$ | $(-1.901 \mathrm{E}-11,1.464 \mathrm{E}-12)$ | 96.19 | $-2.376 \mathrm{E}+00(0.076)$ |
|  | Upper | $9.598 \mathrm{E}-12$ (4.740E-12) | $(-3.543 \mathrm{E}-12,2.278 \mathrm{E}-11)$ | 96.50 | $2.456 \mathrm{E}+00$ (0.070) |
| 6 | Lower | $-9.292 \mathrm{E}-12(3.704 \mathrm{E}-12)$ | (-1.948E-11, $9.938 \mathrm{E}-13)$ | 96.67 | $-2.502 \mathrm{E}+00(0.067)$ |
|  | Upper | $9.130 \mathrm{E}-12$ (4.512E-12) | $(-3.363 \mathrm{E}-12,2.169 \mathrm{E}-11)$ | 96.39 | $2.428 \mathrm{E}+00$ (0.072) |
| 7 | Lower | $-9.162 \mathrm{E}-12(3.095 \mathrm{E}-12)$ | $(-1.767 \mathrm{E}-11,-5.602 \mathrm{E}-13)$ | 97.91 | $-2.956 \mathrm{E}+00(0.042)$ |
|  | Upper | $9.294 \mathrm{E}-12$ (4.597E-12) | $(-3.435 \mathrm{E}-12,2.213 \mathrm{E}-11)$ | 95.38 | $2.202 \mathrm{E}+00$ (0.092) |
| 8 | Lower | $-9.002 \mathrm{E}-12(3.188 \mathrm{E}-12)$ | $(-1.780 \mathrm{E}-11,-1.352 \mathrm{E}-13)$ | 97.61 | $-2.819 \mathrm{E}+00(0.048)$ |
|  | Upper | $9.431 \mathrm{E}-12(4.663 \mathrm{E}-12)$ | $(-3.485 \mathrm{E}-12,2.241 \mathrm{E}-11)$ | 96.47 | $2.447 \mathrm{E}+00(0.071)$ |
| 9 | Lower | $-9.289 \mathrm{E}-12(3.367 \mathrm{E}-12)$ | $(-1.855 \mathrm{E}-11,3.863 \mathrm{E}-14)$ | 97.44 | $-2.754 \mathrm{E}+00(0.051)$ |
|  | Upper | $9.202 \mathrm{E}-12$ (4.551E-12) | $(-3.399 \mathrm{E}-12,2.187 \mathrm{E}-11)$ | 96.41 | $2.433 \mathrm{E}+00$ (0.072) |

Table K-23. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Spline Approximation for $n=30$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | -2.291E-02 (4.187E-03) | $(-3.143 \mathrm{E}-02,-1.438 \mathrm{E}-02)$ | >99.99 | $-5.471 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2931 \\ (<0.0001) \end{gathered}$ |
|  | Upper | 5.887E-03 (4.196E-03) | $(-2.659 \mathrm{E}-03,1.443 \mathrm{E}-02)$ | 91.15 | $1.404 \mathrm{E}+00$ (0.177) | $\begin{gathered} 0.4437 \\ (<0.0001) \end{gathered}$ |
| 1 | Lower | -2.187E-02 (4.193E-03) | $(-3.041 \mathrm{E}-02,-1.333 \mathrm{E}-02)$ | >99.99 | $-5.215 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.3038 \\ (<0.0001) \end{gathered}$ |
|  | Upper | 6.473E-03 (4.192E-03) | $(-2.067 \mathrm{E}-03,1.501 \mathrm{E}-02)$ | 93.38 | $1.544 \mathrm{E}+00(0.132)$ | $\begin{gathered} 0.4362 \\ (<0.0001) \end{gathered}$ |
| 2 | Lower | -2.466E-02 (4.204E-03) | $(-3.322 \mathrm{E}-02,-1.609 \mathrm{E}-02)$ | >99.99 | $-5.865 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2754 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $4.440 \mathrm{E}-03$ (4.219E-03) | $(-4.153 \mathrm{E}-03,1.303 \mathrm{E}-02)$ | 84.98 | $1.052 \mathrm{E}+00(0.300)$ | $\begin{gathered} 0.4599 \\ (<0.0001) \end{gathered}$ |
| 3 | Lower | -2.153E-02 (4.193E-03) | $(-3.007 \mathrm{E}-02,-1.299 \mathrm{E}-02)$ | >99.99 | $-5.134 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.3086 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $7.068 \mathrm{E}-03$ (4.199E-03) | (-1.485E-03, 1.562E-02) | 94.90 | $1.683 \mathrm{E}+00$ (0.102) | $\begin{gathered} 0.4304 \\ (<0.0001) \end{gathered}$ |
| 4 | Lower | -2.376E-02 (4.210E-03) | $(-3.234 \mathrm{E}-02,-1.519 \mathrm{E}-02)$ | >99.99 | $-5.645 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2855 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $5.118 \mathrm{E}-03$ (4.220E-03) | $(-3.478 \mathrm{E}-03,1.371 \mathrm{E}-02)$ | 88.30 | $1.213 \mathrm{E}+00(0.234)$ | $\begin{gathered} 0.4509 \\ (<0.0001) \end{gathered}$ |
| 5 | Lower | -2.205E-02 (4.176E-03) | $(-3.055 \mathrm{E}-02,-1.354 \mathrm{E}-02)$ | >99.99 | $-5.280 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.3017 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $6.709 \mathrm{E}-03$ (4.185E-03) | $(-1.815 \mathrm{E}-03,1.523 \mathrm{E}-02)$ | 94.06 | $1.603 \mathrm{E}+00$ (0.119) | $\begin{gathered} 0.4350 \\ (<0.0001) \end{gathered}$ |
| 6 | Lower | -2.287E-02 (4.194E-03) | $(-3.141 \mathrm{E}-02,-1.433 \mathrm{E}-02)$ | >99.99 | $-5.453 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2930 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $5.868 \mathrm{E}-03$ (4.202E-03) | $(-2.691 \mathrm{E}-03,1.443 \mathrm{E}-02)$ | 91.39 | $1.397 \mathrm{E}+00(0.172)$ | $\begin{gathered} 0.4433 \\ (<0.0001) \end{gathered}$ |
| 7 | Lower | $-2.329 \mathrm{E}-02(4.169 \mathrm{E}-03)$ | $(-3.178 \mathrm{E}-02,-1.480 \mathrm{E}-02)$ | >99.99 | $-5.586 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2886 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $5.612 \mathrm{E}-03$ (4.180E-03) | $(-2.902 \mathrm{E}-03,1.413 \mathrm{E}-02)$ | 90.56 | $1.343 \mathrm{E}+00$ (0.189) | $\begin{gathered} 0.4479 \\ (<0.0001) \end{gathered}$ |
| 8 | Lower | -2.293E-02 (4.164E-03) | $(-3.141 \mathrm{E}-02,-1.444 \mathrm{E}-02)$ | >99.99 | $-5.506 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2917 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $5.990 \mathrm{E}-03$ (4.175E-03) | $(-2.515 \mathrm{E}-03,1.449 \mathrm{E}-02)$ | 91.95 | $1.435 \mathrm{E}+00(0.161)$ | $\begin{gathered} 0.4436 \\ (<0.0001) \end{gathered}$ |
| 9 | Lower | -2.321E-02 (4.179E-03) | $(-3.172 \mathrm{E}-02,-1.470 \mathrm{E}-02)$ | >99.99 | $-5.554 \mathrm{E}+00(<0.001)$ | $\begin{gathered} 0.2899 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $5.709 \mathrm{E}-03$ (4.191E-03) | $(-2.829 \mathrm{E}-03,1.425 \mathrm{E}-02)$ | 90.87 | $1.362 \mathrm{E}+00(0.183)$ | $\begin{gathered} 0.4465 \\ (<0.0001) \end{gathered}$ |

Table K-24. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Spline Approximation for $n=30$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.139 \mathrm{E}-12(2.258 \mathrm{E}-02)$ | (-6.268E-02, 6.268E-02) | 50.00 | $-4.054 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.311 \mathrm{E}-12$ (5.778E-03) | (-1.604E-02, 1.604E-02) | 50.00 | $1.601 \mathrm{E}-09(>0.999)$ |
| 1 | Lower | $-8.679 \mathrm{E}-12$ (2.153E-02) | $(-5.979 \mathrm{E}-02,5.979 \mathrm{E}-02)$ | 50.00 | $-4.031 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.574 \mathrm{E}-12$ (6.374E-03) | $(-1.770 \mathrm{E}-02,1.770 \mathrm{E}-02)$ | 50.00 | $1.502 \mathrm{E}-09$ (>0.999) |
| 2 | Lower | $-9.969 \mathrm{E}-12$ (2.428E-02) | (-6.741E-02, 6.741E-02) | 50.00 | $-4.106 \mathrm{E}-10$ (>0.999) |
|  | Upper | $8.580 \mathrm{E}-12$ (4.373E-03) | $(-1.214 \mathrm{E}-02,1.214 \mathrm{E}-02)$ | 50.00 | $1.962 \mathrm{E}-09$ (>0.999) |
| 3 | Lower | $-8.724 \mathrm{E}-12(2.120 \mathrm{E}-02)$ | $(-5.886 \mathrm{E}-02,5.886 \mathrm{E}-02)$ | 50.00 | $-4.115 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.637 \mathrm{E}-12$ (6.960E-03) | (-1.932E-02, 1.932E-02) | 50.00 | $1.385 \mathrm{E}-09$ (>0.999) |
| 4 | Lower | $-9.773 \mathrm{E}-12(2.340 \mathrm{E}-02)$ | (-6.497E-02, 6.497E-02) | $50.00$ | $-4.176 \mathrm{E}-10(>0.999)$ |
|  | Upper | $8.697 \mathrm{E}-12$ (5.040E-03) | (-1.399E-02, 1.399E-02) | 50.00 | $1.726 \mathrm{E}-09$ (>0.999) |
| 5 | Lower | $-8.789 \mathrm{E}-12(2.171 \mathrm{E}-02)$ | (-6.028E-02, 6.028E-02) | 50.00 | $-4.048 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.608 \mathrm{E}-12(6.607 \mathrm{E}-03)$ | (-1.834E-02, 1.834E-02) | 50.00 | $1.454 \mathrm{E}-09$ (>0.999) |
| 6 | Lower | $-9.269 \mathrm{E}-12(2.252 \mathrm{E}-02)$ | $(-6.253 \mathrm{E}-02,6.253 \mathrm{E}-02)$ | 50.00 | $-4.115 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.148 \mathrm{E}-12$ (5.778E-03) | $(-1.604 \mathrm{E}-02,1.604 \mathrm{E}-02)$ | 50.00 | $1.583 \mathrm{E}-09$ (>0.999) |
| 7 | Lower | $-9.139 \mathrm{E}-12(2.293 \mathrm{E}-02)$ | (-6.367E-02, 6.367E-02) | 50.00 | $-3.985 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.311 \mathrm{E}-12$ (5.526E-03) | $(-1.534 \mathrm{E}-02,1.534 \mathrm{E}-02)$ | 50.00 | $1.685 \mathrm{E}-09$ (>0.999) |
| 8 | Lower | $-8.983 \mathrm{E}-12$ (2.258E-02) | $(-6.268 \mathrm{E}-02,6.268 \mathrm{E}-02)$ | 50.00 | $-3.979 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.446 \mathrm{E}-12$ (5.898E-03) | (-1.638E-02, 1.638E-02) | 50.00 | $1.601 \mathrm{E}-09$ (>0.999) |
| 9 | Lower | $-9.266 \mathrm{E}-12(2.286 \mathrm{E}-02)$ | (-6.346E-02, 6.346E-02) | 50.00 | $-4.054 \mathrm{E}-10$ (>0.999) |
|  | Upper | $9.220 \mathrm{E}-12$ (5.622E-03) | $(-1.561 \mathrm{E}-02,1.561 \mathrm{E}-02)$ | 50.00 | $1.640 \mathrm{E}-09$ (>0.999) |

Table K-25. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat <br> ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | -2.117E-03 (1.104E-03) | (-4.291E-03, 5.623E-05) | 97.02 | $-1.938 \mathrm{E}+00(0.060)$ | $0.2783(<0.0001)$ |
|  | Upper | $1.887 \mathrm{E}-03$ (2.499E-03) | $(-3.035 \mathrm{E}-03,6.809 \mathrm{E}-03)$ | 77.47 | $7.561 \mathrm{E}-01$ (0.451) | 0.1177 (<0.0001) |
| 1 | Lower | $-2.048 \mathrm{E}-03$ (1.261E-03) | (-4.532E-03, 4.361E-04) | 94.72 | $-1.624 \mathrm{E}+00(0.106)$ | 0.2817 (<0.0001) |
|  | Upper | $1.861 \mathrm{E}-03$ (2.731E-03) | (-3.517E-03, 7.238E-03) | 75.19 | $6.814 \mathrm{E}-01$ (0.496) | 0.1178 (<0.0001) |
| 2 | Lower | $-1.977 \mathrm{E}-03(1.088 \mathrm{E}-03)$ | (-4.119E-03, 1.657E-04) | 96.48 | $-1.817 \mathrm{E}+00(0.070)$ | $0.2912(<0.0001)$ |
|  | Upper | $1.985 \mathrm{E}-03(2.514 \mathrm{E}-03)$ | $(-2.966 \mathrm{E}-03,6.936 \mathrm{E}-03)$ | 78.47 | $7.896 \mathrm{E}-01$ (0.431) | 0.1156 (<0.0001) |
| 3 | Lower | $-2.150 \mathrm{E}-03$ (1.203E-03) | (-4.518E-03, 2.186E-04) | 96.25 | $-1.788 \mathrm{E}+00$ (0.075) | $0.2755(<0.0001)$ |
|  | Upper | $1.852 \mathrm{E}-03$ (2.570E-03) | (-3.209E-03, 6.914E-03) | 76.41 | $7.207 \mathrm{E}-01$ (0.472) | 0.1183 (<0.0001) |
| 4 | Lower | $-2.040 \mathrm{E}-03(1.060 \mathrm{E}-03)$ | (-4.128E-03, 4.827E-05) | 97.23 | $-1.924 \mathrm{E}+00(0.055)$ | $0.2909(<0.0001)$ |
|  | Upper | $1.966 \mathrm{E}-03$ (2.439E-03) | (-2.839E-03, 6.770E-03) | 78.94 | $8.058 \mathrm{E}-01$ (0.421) | $0.1161(<0.0001)$ |
| 5 | Lower | $-2.253 \mathrm{E}-03(9.720 \mathrm{E}-04)$ | (-4.167E-03, -3.387E-04) | $98.94$ | $-2.318 \mathrm{E}+00(0.021)$ | $0.2555(<0.0001)$ |
|  | Upper | $1.781 \mathrm{E}-03(2.373 \mathrm{E}-03)$ | $(-2.892 \mathrm{E}-03,6.454 \mathrm{E}-03)$ | 77.32 | $7.507 \mathrm{E}-01$ (0.454) | 0.1199 (<0.0001) |
| 6 | Lower | $-2.140 \mathrm{E}-03(1.178 \mathrm{E}-03)$ | $(-4.461 \mathrm{E}-03,1.806 \mathrm{E}-04)$ | 96.47 | $-1.816 \mathrm{E}+00(0.071)$ | $0.2795(<0.0001)$ |
|  | Upper | $1.884 \mathrm{E}-03$ (2.547E-03) | (-3.132E-03, 6.899E-03) | 76.99 | $7.397 \mathrm{E}-01(0.460)$ | $0.1178(<0.0001)$ |
| 7 | Lower | $-2.182 \mathrm{E}-03$ (1.062E-03) | (-4.274E-03, -9.091E-05) | 97.95 | $-2.055 \mathrm{E}+00(0.041)$ | $0.2715(<0.0001)$ |
|  | Upper | $1.856 \mathrm{E}-03(2.446 \mathrm{E}-03)$ | (-2.962E-03, 6.673E-03) | 77.56 | $7.586 \mathrm{E}-01$ (0.449) | $0.1184(<0.0001)$ |
| 8 | Lower | $-2.084 \mathrm{E}-03(1.164 \mathrm{E}-03)$ | (-4.376E-03, 2.083E-04) | 96.27 | $-1.790 \mathrm{E}+00(0.075)$ | 0.2886 (<0.0001) |
|  | Upper | $1.936 \mathrm{E}-03(2.531 \mathrm{E}-03)$ | $(-3.049 \mathrm{E}-03,6.921 \mathrm{E}-03)$ | 77.75 | $7.650 \mathrm{E}-01$ (0.445) | $0.1167(<0.0001)$ |
| 9 | Lower | $-2.183 \mathrm{E}-03(9.450 \mathrm{E}-04)$ | (-4.044E-03, -3.219E-04) | 98.92 | $-2.310 \mathrm{E}+00(0.022)$ | $0.2706(<0.0001)$ |
|  | Upper | $1.859 \mathrm{E}-03(2.342 \mathrm{E}-03)$ | (-2.754E-03, 6.471E-03) | 78.59 | $7.935 \mathrm{E}-01$ (0.428) | $0.1184(<0.0001)$ |

Table K-26. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | -9.272E-12 (3.191E-14) | $(-9.341 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-2.892 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.190 \mathrm{E}-12$ (2.123E-14) | (9.144E-12, 9.235E-12) | >99.99 | $4.410 \mathrm{E}+02(<0.001)$ |
| 1 | Lower | -9.276E-12 (3.757E-14) | $(-9.356 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-2.469 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.084 \mathrm{E}-12$ (2.505E-14) | $(9.030 \mathrm{E}-12,9.138 \mathrm{E}-12)$ | >99.99 | $3.627 \mathrm{E}+02(<0.001)$ |
| 2 | Lower | $-9.048 \mathrm{E}-12(3.575 \mathrm{E}-14)$ | $(-9.125 \mathrm{E}-12,-8.972 \mathrm{E}-12)$ | >99.99 | $-2.531 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.395 \mathrm{E}-12$ (2.420E-14) | $(9.343 \mathrm{E}-12,9.447 \mathrm{E}-12)$ | >99.99 | $3.882 \mathrm{E}+02(<0.001)$ |
| 3 | Lower | $-9.272 \mathrm{E}-12$ (3.206E-14) | $(-9.341 \mathrm{E}-12,-9.204 \mathrm{E}-12)$ | >99.99 | $-2.892 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.190 \mathrm{E}-12$ (2.123E-14) | (9.144E-12, 9.235E-12) | >99.99 | $4.329 \mathrm{E}+02(<0.001)$ |
| 4 | Lower | $-9.090 \mathrm{E}-12(3.293 \mathrm{E}-14)$ | $(-9.160 \mathrm{E}-12,-9.019 \mathrm{E}-12)$ | >99.99 | $-2.760 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.408 \mathrm{E}-12$ (2.214E-14) | $(9.360 \mathrm{E}-12,9.455 \mathrm{E}-12)$ | >99.99 | $4.250 \mathrm{E}+02(<0.001)$ |
| 5 | Lower | $-9.835 \mathrm{E}-12(2.907 \mathrm{E}-14)$ | $(-9.898 \mathrm{E}-12,-9.773 \mathrm{E}-12)$ | >99.99 | $-3.384 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.769 \mathrm{E}-12$ (1.890E-14) | (8.728E-12, $8.809 \mathrm{E}-12$ ) | >99.99 | $4.640 \mathrm{E}+02(<0.001)$ |
| 6 | Lower | $-9.245 \mathrm{E}-12(3.073 \mathrm{E}-14)$ | $(-9.311 \mathrm{E}-12,-9.179 \mathrm{E}-12)$ | >99.99 | $-3.008 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.280 \mathrm{E}-12(2.034 \mathrm{E}-14)$ | (9.236E-12, 9.324E-12) | >99.99 | $4.563 \mathrm{E}+02(<0.001)$ |
| 7 | Lower | $-9.500 \mathrm{E}-12(2.995 \mathrm{E}-14)$ | $(-9.564 \mathrm{E}-12,-9.436 \mathrm{E}-12)$ | >99.99 | $-3.172 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.091 \mathrm{E}-12(1.978 \mathrm{E}-14)$ | $(9.049 \mathrm{E}-12,9.133 \mathrm{E}-12)$ | >99.99 | $4.595 \mathrm{E}+02(<0.001)$ |
| 8 | Lower | $-9.084 \mathrm{E}-12(3.191 \mathrm{E}-14)$ | $(-9.153 \mathrm{E}-12,-9.016 \mathrm{E}-12)$ | >99.99 | $-2.847 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.431 \mathrm{E}-12(2.138 \mathrm{E}-14)$ | (9.385E-12, $9.477 \mathrm{E}-12)$ | >99.99 | $4.410 \mathrm{E}+02(<0.001)$ |
| 9 | Lower | $-9.631 \mathrm{E}-12(2.947 \mathrm{E}-14)$ | $(-9.694 \mathrm{E}-12,-9.568 \mathrm{E}-12)$ | >99.99 | $-3.269 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.993 \mathrm{E}-12$ (1.938E-14) | (8.951E-12, $9.035 \mathrm{E}-12$ ) | >99.99 | $4.639 \mathrm{E}+02(<0.001)$ |

Table K-27. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat ( $p$-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-2.117 \mathrm{E}-03(2.220 \mathrm{E}-03)$ | (-6.490E-03, 2.256E-03) | 90.12 | $-1.992 \mathrm{E}+00$ (0.198) | $\begin{gathered} 0.3575 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.887 \mathrm{E}-03$ (3.363E-03) | (-4.737E-03, $8.511 \mathrm{E}-03)$ | 74.43 | $8.754 \mathrm{E}-01$ (0.511) | $\begin{gathered} 0.1783 \\ (<0.0001) \end{gathered}$ |
| 1 | Lower | $-2.064 \mathrm{E}-03(2.483 \mathrm{E}-03)$ | (-6.955E-03, 2.827E-03) | 87.91 | $-1.622 \mathrm{E}+00(0.242)$ | $\begin{gathered} 0.3522 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.861 \mathrm{E}-03$ (3.677E-03) | (-5.381E-03, 9.103E-03) | 73.13 | $7.797 \mathrm{E}-01$ (0.537) | $\begin{gathered} 0.1811 \\ (<0.0001) \end{gathered}$ |
| 2 | Lower | $-1.974 \mathrm{E}-03$ (2.245E-03) | (-6.395E-03, 2.446E-03) | 89.20 | $-1.835 \mathrm{E}+00$ (0.216) | $\begin{gathered} 0.3491 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.981 \mathrm{E}-03$ (3.412E-03) | $(-4.740 \mathrm{E}-03,8.701 \mathrm{E}-03)$ | 75.03 | $9.106 \mathrm{E}-01$ (0.499) | $\begin{gathered} 0.1772 \\ (<0.0001) \end{gathered}$ |
| 3 | Lower | $-2.150 \mathrm{E}-03(2.322 \mathrm{E}-03)$ | (-6.723E-03, 2.424E-03) | 89.45 | $-1.806 \mathrm{E}+00$ (0.211) | $\begin{gathered} 0.3610 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.839 \mathrm{E}-03$ (3.436E-03) | $(-4.927 \mathrm{E}-03,8.606 \mathrm{E}-03)$ | 73.86 | 8.262E-01 (0.523) | $\begin{gathered} 0.1804 \\ (<0.0001) \end{gathered}$ |
| 4 | Lower | $-2.043 \mathrm{E}-03(2.137 \mathrm{E}-03)$ | (-6.251E-03, 2.166E-03) | 90.21 | $-1.992 \mathrm{E}+00$ (0.196) | $\begin{gathered} 0.3565 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.980 \mathrm{E}-03$ (3.273E-03) | $(-4.467 \mathrm{E}-03,8.426 \mathrm{E}-03)$ | 75.27 | $9.436 \mathrm{E}-01$ (0.495) | $\begin{gathered} 0.1760 \\ (<0.0001) \end{gathered}$ |
| 5 | Lower | $-2.252 \mathrm{E}-03(2.049 \mathrm{E}-03)$ | (-6.288E-03, 1.784E-03) | 92.22 | $-2.447 \mathrm{E}+00(0.156)$ | $\begin{gathered} 0.3517 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.815 \mathrm{E}-03$ (3.205E-03) | (-4.498E-03, 8.127E-03) | 74.48 | $8.858 \mathrm{E}-01$ (0.510) | $\begin{gathered} 0.1776 \\ (<0.0001) \end{gathered}$ |
| 6 | Lower | $-2.142 \mathrm{E}-03(2.203 \mathrm{E}-03)$ | (-6.481E-03, 2.198E-03) | 89.97 | $-1.871 \mathrm{E}+00$ (0.201) | $\begin{gathered} 0.3638 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.911 \mathrm{E}-03$ (3.339E-03) | $(-4.665 \mathrm{E}-03,8.486 \mathrm{E}-03)$ | 74.33 | $8.645 \mathrm{E}-01$ (0.513) | $\begin{gathered} 0.1756 \\ (<0.0001) \end{gathered}$ |
| 7 | Lower | $-2.170 \mathrm{E}-03(2.211 \mathrm{E}-03)$ | (-6.524E-03, 2.184E-03) | 90.78 | $-2.108 \mathrm{E}+00$ (0.184) | $\begin{gathered} 0.3616 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.838 \mathrm{E}-03$ (3.334E-03) | $(-4.729 \mathrm{E}-03,8.405 \mathrm{E}-03)$ | 74.41 | $8.736 \mathrm{E}-01$ (0.512) | $\begin{gathered} 0.1809 \\ (<0.0001) \end{gathered}$ |
| 8 | Lower | $-2.078 \mathrm{E}-03(2.260 \mathrm{E}-03)$ | (-6.529E-03, 2.373E-03) | 89.44 | $-1.821 \mathrm{E}+00$ (0.211) | $\begin{gathered} 0.3635 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.921 \mathrm{E}-03$ (3.379E-03) | (-4.734E-03, 8.576E-03) | 74.43 | $8.795 \mathrm{E}-01$ (0.511) | $\begin{gathered} 0.1777 \\ (<0.0001) \end{gathered}$ |
| 9 | Lower | $-2.182 \mathrm{E}-03(2.072 \mathrm{E}-03)$ | (-6.264E-03, 1.899E-03) | 91.92 | $-2.421 \mathrm{E}+00(0.162)$ | $\begin{gathered} 0.3579 \\ (<0.0001) \end{gathered}$ |
|  | Upper | $1.842 \mathrm{E}-03$ (3.214E-03) | (-4.488E-03, 8.173E-03) | 74.88 | $9.156 \mathrm{E}-01$ (0.502) | $\begin{gathered} 0.1786 \\ (<0.0001) \end{gathered}$ |

Table K-28. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.272 \mathrm{E}-12(2.882 \mathrm{E}-14)$ | $(-9.336 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-3.103 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.190 \mathrm{E}-12$ (2.123E-14) | $(9.179 \mathrm{E}-12,9.239 \mathrm{E}-12)$ | >99.99 | $4.428 \mathrm{E}+02(<0.001)$ |
| 1 | Lower | $-9.276 \mathrm{E}-12(3.468 \mathrm{E}-14)$ | $(-9.350 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-2.634 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.084 \mathrm{E}-12(2.525 \mathrm{E}-14)$ | (9.073E-12, $9.141 \mathrm{E}-12)$ | >99.99 | $3.616 \mathrm{E}+02(<0.001)$ |
| 2 | Lower | $-9.048 \mathrm{E}-12$ (3.175E-14) | $(-9.116 \mathrm{E}-12,-8.972 \mathrm{E}-12)$ | >99.99 | $-2.827 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.395 \mathrm{E}-12$ (2.420E-14) | (9.387E-12, $9.447 \mathrm{E}-12$ ) | >99.99 | $3.900 \mathrm{E}+02(<0.001)$ |
| 3 | Lower | $-9.272 \mathrm{E}-12(2.980 \mathrm{E}-14)$ | $(-9.336 \mathrm{E}-12,-9.204 \mathrm{E}-12)$ | >99.99 | $-3.090 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.190 \mathrm{E}-12$ (2.123E-14) | (9.179E-12, 9.239E-12) | >99.99 | $4.346 \mathrm{E}+02(<0.001)$ |
| 4 | Lower | $-9.090 \mathrm{E}-12(2.980 \mathrm{E}-14)$ | $(-9.153 \mathrm{E}-12,-9.019 \mathrm{E}-12)$ | >99.99 | $-3.028 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.408 \mathrm{E}-12(2.214 \mathrm{E}-14)$ | $(9.399 \mathrm{E}-12,9.459 \mathrm{E}-12)$ | >99.99 | $4.268 \mathrm{E}+02(<0.001)$ |
| 5 | Lower | $-9.835 \mathrm{E}-12(2.809 \mathrm{E}-14)$ | $(-9.895 \mathrm{E}-12,-9.773 \mathrm{E}-12)$ | >99.99 | $-3.451 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.769 \mathrm{E}-12$ (1.910E-14) | (8.757E-12, 8.817E-12) | >99.99 | $4.608 \mathrm{E}+02(<0.001)$ |
| 6 | Lower | $-9.245 \mathrm{E}-12(2.858 \mathrm{E}-14)$ | $(-9.307 \mathrm{E}-12,-9.179 \mathrm{E}-12)$ | >99.99 | $-3.179 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.280 \mathrm{E}-12(2.054 \mathrm{E}-14)$ | (9.270E-12, 9.327E-12) | >99.99 | $4.536 \mathrm{E}+02(<0.001)$ |
| 7 | Lower | $-9.500 \mathrm{E}-12(2.809 \mathrm{E}-14)$ | $(-9.560 \mathrm{E}-12,-9.436 \mathrm{E}-12)$ | >99.99 | $-3.361 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.091 \mathrm{E}-12(1.978 \mathrm{E}-14)$ | $(9.080 \mathrm{E}-12,9.139 \mathrm{E}-12)$ | >99.99 | $4.612 \mathrm{E}+02(<0.001)$ |
| 8 | Lower | $-9.084 \mathrm{E}-12(2.882 \mathrm{E}-14)$ | $(-9.146 \mathrm{E}-12,-9.016 \mathrm{E}-12)$ | >99.99 | $-3.103 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.431 \mathrm{E}-12$ (2.138E-14) | (9.422E-12, 9.478E-12) | >99.99 | $4.428 \mathrm{E}+02(<0.001)$ |
| 9 | Lower | $-9.631 \mathrm{E}-12(2.784 \mathrm{E}-14)$ | $(-9.690 \mathrm{E}-12,-9.568 \mathrm{E}-12)$ | >99.99 | $-3.438 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.993 \mathrm{E}-12$ (1.938E-14) | (8.982E-12, 9.038E-12) | >99.99 | $4.656 \mathrm{E}+02(<0.001)$ |

Table K-29. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-2.117 \mathrm{E}-03(4.828 \mathrm{E}-05)$ | (-2.212E-03, -2.022E-03) | >99.99 | $-4.388 \mathrm{E}+01(<0.001)$ | $0.2791(<0.0001)$ |
|  | Upper | $1.887 \mathrm{E}-03(4.953 \mathrm{E}-05)$ | (1.789E-03, 1.984E-03) | >99.99 | $3.810 \mathrm{E}+01(<0.001)$ | $0.1177(<0.0001)$ |
| 1 | Lower | $-2.048 \mathrm{E}-03(4.910 \mathrm{E}-05)$ | (-2.145E-03, -1.951E-03) | >99.99 | $-4.171 \mathrm{E}+01(<0.001)$ | 0.2823 (<0.0001) |
|  | Upper | $1.861 \mathrm{E}-03(5.059 \mathrm{E}-05)$ | $(1.761 \mathrm{E}-03,1.960 \mathrm{E}-03)$ | >99.99 | $3.678 \mathrm{E}+01(<0.001)$ | $0.1178(<0.0001)$ |
| 2 | Lower | $-1.977 \mathrm{E}-03(4.929 \mathrm{E}-05)$ | (-2.074E-03, -1.880E-03) | >99.99 | $-4.010 \mathrm{E}+01(<0.001)$ | 0.2917 (<0.0001) |
|  | Upper | $1.985 \mathrm{E}-03$ (5.049E-05) | (1.886E-03, 2.084E-03) | >99.99 | $3.932 \mathrm{E}+01(<0.001)$ | $0.1156(<0.0001)$ |
| 3 | Lower | $-2.150 \mathrm{E}-03$ (4.745E-05) | (-2.243E-03, -2.056E-03) | >99.99 | $-4.530 \mathrm{E}+01(<0.001)$ | 0.2763 (<0.0001) |
|  | Upper | $1.852 \mathrm{E}-03$ (4.880E-05) | (1.756E-03, 1.949E-03) | >99.99 | $3.796 \mathrm{E}+01(<0.001)$ | 0.1183 (<0.0001) |
| 4 | Lower | $-2.040 \mathrm{E}-03(4.813 \mathrm{E}-05)$ | $(-2.135 \mathrm{E}-03,-1.945 \mathrm{E}-03)$ | >99.99 | $-4.238 \mathrm{E}+01(<0.001)$ | 0.2915 (<0.0001) |
|  | Upper | $1.966 \mathrm{E}-03$ (4.929E-05) | $(1.869 \mathrm{E}-03,2.063 \mathrm{E}-03)$ | >99.99 | $3.988 \mathrm{E}+01(<0.001)$ | $0.1161(<0.0001)$ |
| 5 | Lower | $-2.253 \mathrm{E}-03(4.899 \mathrm{E}-05)$ | $(-2.350 \mathrm{E}-03,-2.157 \mathrm{E}-03)$ | >99.99 | $-4.599 \mathrm{E}+01(<0.001)$ | $0.2564(<0.0001)$ |
|  | Upper | $1.781 \mathrm{E}-03(5.015 \mathrm{E}-05)$ | $(1.683 \mathrm{E}-03,1.880 \mathrm{E}-03)$ | >99.99 | $3.552 \mathrm{E}+01(<0.001)$ | $0.1199(<0.0001)$ |
| 6 | Lower | $-2.140 \mathrm{E}-03(4.741 \mathrm{E}-05)$ | $(-2.233 \mathrm{E}-03,-2.047 \mathrm{E}-03)$ | >99.99 | $-4.514 \mathrm{E}+01(<0.001)$ | $0.2803(<0.0001)$ |
|  | Upper | $1.884 \mathrm{E}-03(4.873 \mathrm{E}-05)$ | (1.788E-03, 1.980E-03) | >99.99 | $3.866 \mathrm{E}+01(<0.001)$ | $0.1178(<0.0001)$ |
| 7 | Lower | $-2.182 \mathrm{E}-03$ (4.808E-05) | (-2.277E-03, -2.088E-03) | >99.99 | $-4.539 \mathrm{E}+01(<0.001)$ | 0.2723 (<0.0001) |
|  | Upper | $1.856 \mathrm{E}-03(4.930 \mathrm{E}-05)$ | (1.759E-03, 1.953E-03) | >99.99 | $3.764 \mathrm{E}+01(<0.001)$ | $0.1185(<0.0001)$ |
| 8 | Lower | $-2.084 \mathrm{E}-03(4.738 \mathrm{E}-05)$ | $(-2.177 \mathrm{E}-03,-1.991 \mathrm{E}-03)$ | >99.99 | $-4.398 \mathrm{E}+01(<0.001)$ | 0.2893 (<0.0001) |
|  | Upper | $1.936 \mathrm{E}-03$ (4.866E-05) | (1.840E-03, 2.032E-03) | >99.99 | $3.979 \mathrm{E}+01(<0.001)$ | $0.1168(<0.0001)$ |
| 9 | Lower | $-2.183 \mathrm{E}-03(4.863 \mathrm{E}-05)$ | $(-2.279 \mathrm{E}-03,-2.087 \mathrm{E}-03)$ | >99.99 | $-4.489 \mathrm{E}+01(<0.001)$ | 0.2715 (<0.0001) |
|  | Upper | $1.859 \mathrm{E}-03$ (4.974E-05) | (1.761E-03, 1.956E-03) | >99.99 | $3.736 \mathrm{E}+01(<0.001)$ | $0.1184(<0.0001)$ |

Table K-30. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Linear Trapezoid Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.272 \mathrm{E}-12$ (2.136E-03) | (-4.581E-03, 4.581E-03) | 50.00 | $-4.374 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.208 \mathrm{E}-12$ (1.857E-03) | (-3.983E-03, 3.983E-03) | 50.00 | $4.904 \mathrm{E}-09$ (>0.999) |
| 1 | Lower | $-9.276 \mathrm{E}-12(2.044 \mathrm{E}-03)$ | (-4.384E-03, 4.384E-03) | 50.00 | $-4.538 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.106 \mathrm{E}-12$ (1.857E-03) | (-3.983E-03, 3.983E-03) | 50.00 | $4.904 \mathrm{E}-09$ ( $>0.999$ ) |
| 2 | Lower | $-9.048 \mathrm{E}-12$ (1.973E-03) | (-4.231E-03, 4.231E-03) | 50.00 | -4.586E-09 (>0.999) |
|  | Upper | $9.417 \mathrm{E}-12$ (1.981E-03) | (-4.249E-03, 4.249E-03) | 50.00 | $4.753 \mathrm{E}-09$ (>0.999) |
| 3 | Lower | $-9.272 \mathrm{E}-12$ (2.145E-03) | (-4.601E-03, 4.601E-03) | 50.00 | $-4.322 \mathrm{E}-09$ (>0.999) |
|  | Upper | $9.208 \mathrm{E}-12$ (1.849E-03) | (-3.965E-03, 3.965E-03) | 50.00 | $4.981 \mathrm{E}-09$ (>0.999) |
| 4 | Lower | $-9.090 \mathrm{E}-12$ (2.036E-03) | (-4.366E-03, 4.366E-03) | 50.00 | $-4.465 \mathrm{E}-09$ (>0.999) |
|  | Upper | $9.428 \mathrm{E}-12$ (1.962E-03) | (-4.208E-03, 4.208E-03) | 50.00 | $4.805 \mathrm{E}-09$ (>0.999) |
| 5 | Lower | $-9.835 \mathrm{E}-12$ (2.249E-03) | (-4.823E-03, 4.823E-03) | 50.00 | $-4.374 \mathrm{E}-09$ (>0.999) |
|  | Upper | $8.784 \mathrm{E}-12$ (1.778E-03) | $(-3.813 \mathrm{E}-03,3.813 \mathrm{E}-03)$ | 50.00 | $4.941 \mathrm{E}-09$ (>0.999) |
| 6 | Lower | $-9.245 \mathrm{E}-12$ (2.136E-03) | (-4.581E-03, 4.581E-03) | 50.00 | $-4.329 \mathrm{E}-09$ (>0.999) |
|  | Upper | $9.298 \mathrm{E}-12$ (1.880E-03) | (-4.032E-03, 4.032E-03) | 50.00 | $4.945 \mathrm{E}-09$ (>0.999) |
| 7 | Lower | $-9.500 \mathrm{E}-12(2.178 \mathrm{E}-03)$ | (-4.672E-03, 4.672E-03) | 50.00 | $-4.362 \mathrm{E}-09$ (>0.999) |
|  | Upper | $9.107 \mathrm{E}-12$ (1.852E-03) | (-3.972E-03, 3.972E-03) | 50.00 | $4.917 \mathrm{E}-09$ (>0.999) |
| 8 | Lower | $-9.084 \mathrm{E}-12(2.080 \mathrm{E}-03)$ | (-4.461E-03, 4.461E-03) | 50.00 | $-4.368 \mathrm{E}-09$ (>0.999) |
|  | Upper | $9.450 \mathrm{E}-12$ (1.933E-03) | $(-4.145 \mathrm{E}-03,4.145 \mathrm{E}-03)$ | 50.00 | $4.890 \mathrm{E}-09$ (>0.999) |
| 9 | Lower | $-9.631 \mathrm{E}-12$ (2.179E-03) | (-4.673E-03, 4.673E-03) | 50.00 | $-4.421 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.009 \mathrm{E}-12$ (1.855E-03) | (-3.978E-03, 3.978E-03) | 50.00 | $4.857 \mathrm{E}-09$ (>0.999) |

Table K-31. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Spline Approximation for $n=250$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-5.733 \mathrm{E}-03(1.536 \mathrm{E}-02)$ | $(-3.598 \mathrm{E}-02,2.451 \mathrm{E}-02)$ | 64.54 | -3.734E-01 (0.709) | $0.0793(<0.0001)$ |
|  | Upper | $-1.274 \mathrm{E}-03$ (1.300E-02) | (-2.688E-02, 2.433E-02) | 53.90 | -9.801E-02 (0.922) | 0.0916 (<0.0001) |
| 1 | Lower | $-5.684 \mathrm{E}-03(1.553 \mathrm{E}-02)$ | $(-3.627 \mathrm{E}-02,2.490 \mathrm{E}-02)$ | 64.27 | -3.660E-01 (0.715) | $0.0794(<0.0001)$ |
|  | Upper | $-1.318 \mathrm{E}-03$ (1.356E-02) | $(-2.802 \mathrm{E}-02,2.539 \mathrm{E}-02)$ | 53.87 | $-9.720 \mathrm{E}-02$ (0.923) | 0.0913 (<0.0001) |
| 2 | Lower | $-5.598 \mathrm{E}-03(1.547 \mathrm{E}-02)$ | $(-3.607 \mathrm{E}-02,2.488 \mathrm{E}-02)$ | 64.11 | -3.618E-01 (0.718) | 0.0800 (<0.0001) |
|  | Upper | $-1.180 \mathrm{E}-03(1.327 \mathrm{E}-02)$ | (-2.731E-02, 2.495E-02) | 53.54 | -8.894E-02 (0.929) | $0.0913(<0.0001)$ |
| 3 | Lower | $-5.769 \mathrm{E}-03(1.544 \mathrm{E}-02)$ | (-3.618E-02, 2.464E-02) | 64.55 | -3.736E-01 (0.709) | 0.0790 (<0.0001) |
|  | Upper | $-1.311 \mathrm{E}-03$ (1.298E-02) | (-2.687E-02, 2.425E-02) | 54.02 | -1.010E-01 (0.920) | 0.0916 (<0.0001) |
| 4 | Lower | $-5.654 \mathrm{E}-03(1.544 \mathrm{E}-02)$ | (-3.607E-02, 2.476E-02) | 64.27 | -3.661E-01 (0.715) | $0.0797(<0.0001)$ |
|  | Upper | $-1.193 \mathrm{E}-03(1.300 \mathrm{E}-02)$ | (-2.680E-02, 2.441E-02) | 53.65 | -9.172E-02 (0.927) | 0.0915 (<0.0001) |
| 5 | Lower | $-5.866 \mathrm{E}-03$ (1.516E-02) | (-3.572E-02, 2.399E-02) | 65.04 | -3.869E-01 (0.699) | 0.0786 (<0.0001) |
|  | Upper | $-1.376 \mathrm{E}-03$ (1.281E-02) | (-2.661E-02, 2.386E-02) | 54.27 | -1.074E-01 (0.915) | $0.0918(<0.0001)$ |
| 6 | Lower | $-5.752 \mathrm{E}-03$ (1.536E-02) | (-3.600E-02, 2.450E-02) | 64.58 | -3.745E-01 (0.708) | $0.0792(<0.0001)$ |
|  | Upper | $-1.273 \mathrm{E}-03$ (1.288E-02) | $(-2.664 \mathrm{E}-02,2.410 \mathrm{E}-02)$ | 53.93 | $-9.883 \mathrm{E}-02$ (0.921) | 0.0917 (<0.0001) |
| 7 | Lower | $-5.792 \mathrm{E}-03(1.523 \mathrm{E}-02)$ | $(-3.579 \mathrm{E}-02,2.421 \mathrm{E}-02)$ | 64.79 | -3.802E-01 (0.704) | $0.0790(<0.0001)$ |
|  | Upper | $-1.299 \mathrm{E}-03(1.281 \mathrm{E}-02)$ | (-2.653E-02, 2.393E-02) | 54.03 | -1.014E-01 (0.919) | $0.0918(<0.0001)$ |
| 8 | Lower | $-5.695 \mathrm{E}-03(1.541 \mathrm{E}-02)$ | (-3.604E-02, 2.465E-02) | 64.40 | $-3.697 \mathrm{E}-01(0.712)$ | $0.0795(<0.0001)$ |
|  | Upper | $-1.220 \mathrm{E}-03$ (1.292E-02) | (-2.666E-02, 2.422E-02) | 53.76 | $-9.444 \mathrm{E}-02$ (0.925) | 0.0916 (<0.0001) |
| 9 | Lower | $-5.791 \mathrm{E}-03(1.517 \mathrm{E}-02)$ | $(-3.567 \mathrm{E}-02,2.409 \mathrm{E}-02)$ | 64.85 | $-3.817 \mathrm{E}-01(0.703)$ | $0.0791(<0.0001)$ |
|  | Upper | $-1.295 \mathrm{E}-03(1.280 \mathrm{E}-02)$ | $(-2.650 \mathrm{E}-02,2.391 \mathrm{E}-02)$ | 54.02 | $-1.012 \mathrm{E}-01(0.920)$ | $0.0918(<0.0001)$ |

Table K-32. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Partial Batch Estimation to Spline Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.273 \mathrm{E}-12(3.144 \mathrm{E}-14)$ | $(-9.342 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-2.889 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.189 \mathrm{E}-12$ (1.847E-14) | (9.150E-12, 9.229E-12) | >99.99 | $4.974 \mathrm{E}+02(<0.001)$ |
| 1 | Lower | $-9.276 \mathrm{E}-12(3.759 \mathrm{E}-14)$ | $(-9.356 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-2.467 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.084 \mathrm{E}-12$ (2.229E-14) | (9.036E-12, $9.132 \mathrm{E}-12$ ) | >99.99 | $4.075 \mathrm{E}+02(<0.001)$ |
| 2 | Lower | $-9.047 \mathrm{E}-12(3.504 \mathrm{E}-14)$ | $(-9.123 \mathrm{E}-12,-8.972 \mathrm{E}-12)$ | >99.99 | $-2.582 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.396 \mathrm{E}-12$ (2.198E-14) | (9.348E-12, 9.443E-12) | >99.99 | $4.274 \mathrm{E}+02(<0.001)$ |
| 3 | Lower | $-9.273 \mathrm{E}-12(3.231 \mathrm{E}-14)$ | $(-9.342 \mathrm{E}-12,-9.203 \mathrm{E}-12)$ | >99.99 | $-2.870 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.189 \mathrm{E}-12$ (1.847E-14) | $(9.150 \mathrm{E}-12,9.229 \mathrm{E}-12)$ | >99.99 | $4.974 \mathrm{E}+02(<0.001)$ |
| 4 | Lower | $-9.090 \mathrm{E}-12(3.291 \mathrm{E}-14)$ | $(-9.160 \mathrm{E}-12,-9.019 \mathrm{E}-12)$ | >99.99 | $-2.762 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.407 \mathrm{E}-12$ (2.023E-14) | (9.364E-12, $9.451 \mathrm{E}-12$ ) | >99.99 | $4.651 \mathrm{E}+02(<0.001)$ |
| 5 | Lower | $-9.836 \mathrm{E}-12(2.953 \mathrm{E}-14)$ | $(-9.899 \mathrm{E}-12,-9.772 \mathrm{E}-12)$ | >99.99 | $-3.331 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.768 \mathrm{E}-12(1.541 \mathrm{E}-14)$ | (8.735E-12, 8.801E-12) | >99.99 | $5.690 \mathrm{E}+02(<0.001)$ |
| 6 | Lower | $-9.246 \mathrm{E}-12(3.120 \mathrm{E}-14)$ | $(-9.313 \mathrm{E}-12,-9.179 \mathrm{E}-12)$ | >99.99 | $-2.964 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.280 \mathrm{E}-12(1.823 \mathrm{E}-14)$ | $(9.241 \mathrm{E}-12,9.319 \mathrm{E}-12)$ | >99.99 | $5.090 \mathrm{E}+02(<0.001)$ |
| 7 | Lower | $-9.500 \mathrm{E}-12(2.949 \mathrm{E}-14)$ | $(-9.563 \mathrm{E}-12,-9.436 \mathrm{E}-12)$ | >99.99 | $-3.222 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.091 \mathrm{E}-12(1.668 \mathrm{E}-14)$ | (9.055E-12, 9.127E-12) | >99.99 | $5.452 \mathrm{E}+02(<0.001)$ |
| 8 | Lower | $-9.083 \mathrm{E}-12(3.144 \mathrm{E}-14)$ | $(-9.151 \mathrm{E}-12,-9.016 \mathrm{E}-12)$ | >99.99 | $-2.889 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.432 \mathrm{E}-12(1.907 \mathrm{E}-14)$ | $(9.391 \mathrm{E}-12,9.472 \mathrm{E}-12)$ | >99.99 | $4.945 \mathrm{E}+02(<0.001)$ |
| 9 | Lower | $-9.631 \mathrm{E}-12(2.944 \mathrm{E}-14)$ | $(-9.694 \mathrm{E}-12,-9.568 \mathrm{E}-12)$ | >99.99 | $-3.271 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.993 \mathrm{E}-12$ (1.599E-14) | (8.959E-12, $9.027 \mathrm{E}-12$ ) | >99.99 | $5.625 \mathrm{E}+02(<0.001)$ |

Table K-33. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Spline Approximation for $n=250$ Patients

| AE | Effect | Paired Difference Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-5.731 \mathrm{E}-03$ (1.337E-02) | (-3.207E-02, 2.061E-02) | 72.62 | -7.370E-01 (0.548) | 0.1465 (<0.0001) |
|  | Upper | $-1.284 \mathrm{E}-03$ (9.898E-03) | $(-2.078 \mathrm{E}-02,1.821 \mathrm{E}-02)$ | 78.14 | $-6.444 \mathrm{E}-02(0.437)$ | $0.1512(<0.0001)$ |
| 1 | Lower | $-5.657 \mathrm{E}-03(1.285 \mathrm{E}-02)$ | (-3.096E-02, 1.965E-02) | 73.90 | $-8.720 \mathrm{E}-01(0.522)$ | $0.1450(<0.0001)$ |
|  | Upper | $-1.285 \mathrm{E}-03$ (1.028E-02) | $(-2.153 \mathrm{E}-02,1.896 \mathrm{E}-02)$ | 78.10 | $2.973 \mathrm{E}-01$ (0.438) | 0.1496 (<0.0001) |
| 2 | Lower | $-5.574 \mathrm{E}-03(1.338 \mathrm{E}-02)$ | (-3.193E-02, 2.078E-02) | 72.91 | $-7.507 \mathrm{E}-01(0.542)$ | 0.1478 (<0.0001) |
|  | Upper | $-1.139 \mathrm{E}-03$ (1.023E-02) | $(-2.128 \mathrm{E}-02,1.900 \mathrm{E}-02)$ | 77.74 | $5.224 \mathrm{E}-01(0.445)$ | $0.1524(<0.0001)$ |
| 3 | Lower | $-5.891 \mathrm{E}-03(1.364 \mathrm{E}-02)$ | (-3.275E-02, 2.096E-02) | 72.60 | $-7.252 \mathrm{E}-01$ (0.548) | $0.1482(<0.0001)$ |
|  | Upper | $-1.439 \mathrm{E}-03(9.934 \mathrm{E}-03)$ | $(-2.100 \mathrm{E}-02,1.813 \mathrm{E}-02)$ | 78.18 | -3.116E-01 (0.436) | 0.1530 (<0.0001) |
| 4 | Lower | $-5.571 \mathrm{E}-03(1.355 \mathrm{E}-02)$ | (-3.225E-02, 2.111E-02) | 72.16 | -6.958E-01 (0.557) | $0.1471(<0.0001)$ |
|  | Upper | $-1.164 \mathrm{E}-03(9.875 \mathrm{E}-03)$ | $(-2.061 \mathrm{E}-02,1.828 \mathrm{E}-02)$ | 78.04 | -2.145E-01 (0.439) | 0.1517 (<0.0001) |
| 5 | Lower | $-5.830 \mathrm{E}-03(1.331 \mathrm{E}-02)$ | (-3.205E-02, 2.039E-02) | 72.58 | $-7.409 \mathrm{E}-01(0.548)$ | 0.1473 (<0.0001) |
|  | Upper | $-1.349 \mathrm{E}-03$ (9.815E-03) | $(-2.068 \mathrm{E}-02,1.798 \mathrm{E}-02)$ | 78.18 | $6.992 \mathrm{E}-02$ (0.436) | $0.1521(<0.0001)$ |
| 6 | Lower | $-5.761 \mathrm{E}-03(1.352 \mathrm{E}-02)$ | $(-3.240 \mathrm{E}-02,2.087 \mathrm{E}-02)$ | 72.31 | $-6.979 \mathrm{E}-01(0.554)$ | $0.1459(<0.0001)$ |
|  | Upper | $-1.315 \mathrm{E}-03(9.792 \mathrm{E}-03)$ | (-2.060E-02, 1.797E-02) | 78.20 | $-3.842 \mathrm{E}-01(0.436)$ | $0.1506(<0.0001)$ |
| 7 | Lower | $-5.799 \mathrm{E}-03(1.347 \mathrm{E}-02)$ | (-3.233E-02, 2.073E-02) | 72.36 | $-7.288 \mathrm{E}-01(0.553)$ | $0.1481(<0.0001)$ |
|  | Upper | $-1.361 \mathrm{E}-03(9.823 \mathrm{E}-03)$ | $(-2.071 \mathrm{E}-02,1.798 \mathrm{E}-02)$ | 78.16 | -2.457E-01 (0.437) | $0.1527(<0.0001)$ |
| 8 | Lower | $-5.678 \mathrm{E}-03(1.352 \mathrm{E}-02)$ | (-3.231E-02, 2.095E-02) | 72.08 | $-6.792 \mathrm{E}-01(0.558)$ | $0.1450(<0.0001)$ |
|  | Upper | $-1.206 \mathrm{E}-03(9.768 \mathrm{E}-03)$ | $(-2.044 \mathrm{E}-02,1.803 \mathrm{E}-02)$ | 78.07 | -3.653E-01 (0.439) | 0.1496 (<0.0001) |
| 9 | Lower | $-5.822 \mathrm{E}-03(1.313 \mathrm{E}-02)$ | $(-3.167 \mathrm{E}-02,2.003 \mathrm{E}-02)$ | 72.69 | $-7.424 \mathrm{E}-01(0.546)$ | $0.1442(<0.0001)$ |
|  | Upper | $-1.293 \mathrm{E}-03(9.567 \mathrm{E}-03)$ | (-2.013E-02, 1.755E-02) | 78.60 | $5.161 \mathrm{E}-02$ (0.428) | 0.1486 (<0.0001) |

Table K-34. Nonparametric Equivalence Analyses for $\Delta=10 \%$ Using Bootstrap Estimation to Spline Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-9.273 \mathrm{E}-12(2.882 \mathrm{E}-14)$ | $(-9.337 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-3.130 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.189 \mathrm{E}-12$ (2.096E-14) | $(9.150 \mathrm{E}-12,9.239 \mathrm{E}-12)$ | >99.99 | $4.444 \mathrm{E}+02(<0.001)$ |
| 1 | Lower | $-9.276 \mathrm{E}-12(3.468 \mathrm{E}-14)$ | $(-9.350 \mathrm{E}-12,-9.195 \mathrm{E}-12)$ | >99.99 | $-2.651 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.084 \mathrm{E}-12$ (2.522E-14) | (9.037E-12, $9.141 \mathrm{E}-12$ ) | >99.99 | $3.619 \mathrm{E}+02(<0.001)$ |
| 2 | Lower | $-9.047 \mathrm{E}-12$ (3.175E-14) | $(-9.114 \mathrm{E}-12,-8.972 \mathrm{E}-12)$ | >99.99 | $-2.825 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.396 \mathrm{E}-12$ (2.398E-14) | $(9.349 \mathrm{E}-12,9.448 \mathrm{E}-12)$ | >99.99 | $3.936 \mathrm{E}+02(<0.001)$ |
| 3 | Lower | $-9.273 \mathrm{E}-12(3.004 \mathrm{E}-14)$ | $(-9.337 \mathrm{E}-12,-9.203 \mathrm{E}-12)$ | >99.99 | $-3.087 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.189 \mathrm{E}-12$ (2.163E-14) | $(9.150 \mathrm{E}-12,9.239 \mathrm{E}-12)$ | >99.99 | $4.266 \mathrm{E}+02(<0.001)$ |
| 4 | Lower | $-9.090 \mathrm{E}-12(2.931 \mathrm{E}-14)$ | (-9.152E-12, -9.019E-12) | >99.99 | $-3.101 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.407 \mathrm{E}-12(2.216 \mathrm{E}-14)$ | (9.365E-12, 9.457E-12) | >99.99 | $4.263 \mathrm{E}+02(<0.001)$ |
| 5 | Lower | $-9.836 \mathrm{E}-12(2.784 \mathrm{E}-14)$ | $(-9.895 \mathrm{E}-12,-9.774 \mathrm{E}-12)$ | >99.99 | $-3.451 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.768 \mathrm{E}-12$ (1.932E-14) | (8.735E-12, 8.815E-12) | >99.99 | $4.555 \mathrm{E}+02(<0.001)$ |
| 6 | Lower | $-9.246 \mathrm{E}-12(2.882 \mathrm{E}-14)$ | $(-9.308 \mathrm{E}-12,-9.179 \mathrm{E}-12)$ | >99.99 | $-3.186 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.280 \mathrm{E}-12(2.096 \mathrm{E}-14)$ | (9.242E-12, 9.328E-12) | >99.99 | $4.444 \mathrm{E}+02(<0.001)$ |
| 7 | Lower | $-9.500 \mathrm{E}-12(2.809 \mathrm{E}-14)$ | $(-9.559 \mathrm{E}-12,-9.436 \mathrm{E}-12)$ | >99.99 | $-3.332 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.091 \mathrm{E}-12$ (1.976E-14) | (9.056E-12, 9.137E-12) | >99.99 | $4.620 \mathrm{E}+02(<0.001)$ |
| 8 | Lower | $-9.083 \mathrm{E}-12(2.882 \mathrm{E}-14)$ | $(-9.145 \mathrm{E}-12,-9.016 \mathrm{E}-12)$ | >99.99 | $-3.130 \mathrm{E}+02(<0.001)$ |
|  | Upper | $9.432 \mathrm{E}-12$ (2.096E-14) | $(9.391 \mathrm{E}-12,9.478 \mathrm{E}-12)$ | >99.99 | $4.517 \mathrm{E}+02(<0.001)$ |
| 9 | Lower | $-9.631 \mathrm{E}-12(2.760 \mathrm{E}-14)$ | $(-9.689 \mathrm{E}-12,-9.568 \mathrm{E}-12)$ | >99.99 | $-3.438 \mathrm{E}+02(<0.001)$ |
|  | Upper | $8.993 \mathrm{E}-12$ (1.941E-14) | (8.959E-12, $9.039 \mathrm{E}-12$ ) | >99.99 | $4.650 \mathrm{E}+02(<0.001)$ |

Table K-35. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Spline Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Mean (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) | S-W Stat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | $-5.733 \mathrm{E}-03$ (5.366E-04) | (-6.790E-03, -4.677E-03) | >99.99 | $-1.069 \mathrm{E}+01(<0.001)$ | $0.0793(<0.0001)$ |
|  | Upper | $-1.274 \mathrm{E}-03$ (5.080E-04) | (-2.274E-03, -2.733E-04) | 99.33 | $-2.508 \mathrm{E}+00(0.013)$ | $0.0915(<0.0001)$ |
| 1 | Lower | $-5.684 \mathrm{E}-03(5.434 \mathrm{E}-04)$ | (-6.754E-03, -4.614E-03) | >99.99 | $-1.046 \mathrm{E}+01(<0.001)$ | $0.0794(<0.0001)$ |
|  | Upper | $-1.318 \mathrm{E}-03$ (5.183E-04) | (-2.339E-03, -2.973E-04) | 99.42 | $-2.543 \mathrm{E}+00(0.012)$ | $0.0912(<0.0001)$ |
| 2 | Lower | $-5.598 \mathrm{E}-03(5.410 \mathrm{E}-04)$ | (-6.664E-03, -4.533E-03) | >99.99 | $-1.035 \mathrm{E}+01(<0.001)$ | $0.0800(<0.0001)$ |
|  | Upper | $-1.180 \mathrm{E}-03$ (5.136E-04) | $(-2.192 \mathrm{E}-03,-1.687 \mathrm{E}-04)$ | 98.88 | $-2.298 \mathrm{E}+00(0.022)$ | $0.0912(<0.0001)$ |
| 3 | Lower | $-5.769 \mathrm{E}-03(5.368 \mathrm{E}-04)$ | (-6.826E-03, -4.711E-03) | >99.99 | $-1.075 \mathrm{E}+01(<0.001)$ | 0.0790 (<0.0001) |
|  | Upper | $-1.311 \mathrm{E}-03(5.067 \mathrm{E}-04)$ | (-2.309E-03, -3.129E-04) | 99.49 | $-2.587 \mathrm{E}+00(0.010)$ | $0.0916(<0.0001)$ |
| 4 | Lower | $-5.654 \mathrm{E}-03(5.380 \mathrm{E}-04)$ | $(-6.713 \mathrm{E}-03,-4.594 \mathrm{E}-03)$ | >99.99 | $-1.051 \mathrm{E}+01(<0.001)$ | $0.0798(<0.0001)$ |
|  | Upper | $-1.193 \mathrm{E}-03(5.081 \mathrm{E}-04)$ | $(-2.193 \mathrm{E}-03,-1.918 \mathrm{E}-04)$ | 99.01 | $-2.347 \mathrm{E}+00(0.020)$ | 0.0914 (<0.0001) |
| 5 | Lower | $-5.866 \mathrm{E}-03(5.327 \mathrm{E}-04)$ | (-6.915E-03, -4.816E-03) | >99.99 | $-1.101 \mathrm{E}+01(<0.001)$ | $0.0787(<0.0001)$ |
|  | Upper | $-1.376 \mathrm{E}-03$ (5.050E-04) | (-2.371E-03, -3.815E-04) | 99.66 | $-2.725 \mathrm{E}+00(0.007)$ | 0.0917 (<0.0001) |
| 6 | Lower | $-5.752 \mathrm{E}-03(5.349 \mathrm{E}-04)$ | (-6.806E-03, -4.699E-03) | >99.99 | $-1.075 \mathrm{E}+01(<0.001)$ | $0.0792(<0.0001)$ |
|  | Upper | $-1.273 \mathrm{E}-03(5.050 \mathrm{E}-04)$ | (-2.268E-03, -2.785E-04) | 99.38 | $-2.521 \mathrm{E}+00(0.012)$ | $0.0916(<0.0001)$ |
| 7 | Lower | $-5.792 \mathrm{E}-03$ (5.332E-04) | (-6.842E-03, -4.742E-03) | >99.99 | $-1.086 \mathrm{E}+01(<0.001)$ | $0.0791(<0.0001)$ |
|  | Upper | $-1.299 \mathrm{E}-03(5.044 \mathrm{E}-04)$ | $(-2.292 \mathrm{E}-03,-3.051 \mathrm{E}-04)$ | 99.47 | $-2.574 \mathrm{E}+00(0.011)$ | 0.0917 (<0.0001) |
| 8 | Lower | $-5.695 \mathrm{E}-03(5.360 \mathrm{E}-04)$ | (-6.751E-03, -4.640E-03) | >99.99 | $-1.063 \mathrm{E}+01(<0.001)$ | 0.0796 (<0.0001) |
|  | Upper | $-1.220 \mathrm{E}-03(5.058 \mathrm{E}-04)$ | $(-2.216 \mathrm{E}-03,-2.236 \mathrm{E}-04)$ | 99.17 | $-2.412 \mathrm{E}+00(0.017)$ | 0.0915 (<0.0001) |
| 9 | Lower | $-5.791 \mathrm{E}-03(5.329 \mathrm{E}-04)$ | (-6.841E-03, -4.742E-03) | >99.99 | $-1.087 \mathrm{E}+01(<0.001)$ | $0.0792(<0.0001)$ |
|  | Upper | $-1.295 \mathrm{E}-03$ (5.049E-04) | $(-2.289 \mathrm{E}-03,-3.002 \mathrm{E}-04)$ | 99.45 | $-2.564 \mathrm{E}+00(0.011)$ | $0.0917(<0.0001)$ |

Table K-36. Parametric Equivalence Analyses for $\Delta=10 \%$ Using Jackknife Estimation to Spline Approximation for $n=250$ Patients

| AE | Effect | Paired Difference <br> Median (SE) | 95\% CI on Paired Difference | Coverage (\%) | Tstat (p-value) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Overall | Lower | -9.273E-12 (5.741E-03) | $(-1.231 \mathrm{E}-02,1.231 \mathrm{E}-02)$ | 50.00 | -1.619E-09 ( $>0.999$ ) |
|  | Upper | $9.189 \mathrm{E}-12$ (1.292E-03) | $(-2.771 \mathrm{E}-03,2.771 \mathrm{E}-03)$ | 50.00 | $7.024 \mathrm{E}-09$ (>0.999) |
| 1 | Lower | $-9.276 \mathrm{E}-12$ (5.673E-03) | $(-1.217 \mathrm{E}-02,1.217 \mathrm{E}-02)$ | 50.00 | $-1.635 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.084 \mathrm{E}-12$ (1.315E-03) | $(-2.821 \mathrm{E}-03,2.821 \mathrm{E}-03)$ | 50.00 | $6.906 \mathrm{E}-09$ (>0.999) |
| 2 | Lower | $-9.047 \mathrm{E}-12$ (5.587E-03) | (-1.198E-02, 1.198E-02) | 50.00 | $-1.619 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.396 \mathrm{E}-12$ (1.178E-03) | (-2.526E-03, 2.526E-03) | 50.00 | $7.977 \mathrm{E}-09$ (>0.999) |
| 3 | Lower | $-9.273 \mathrm{E}-12(5.757 \mathrm{E}-03)$ | $(-1.235 \mathrm{E}-02,1.235 \mathrm{E}-02)$ | 50.00 | $-1.611 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.189 \mathrm{E}-12$ (1.308E-03) | (-2.806E-03, 2.806E-03) | 50.00 | $7.024 \mathrm{E}-09$ (>0.999) |
| 4 | Lower | $-9.090 \mathrm{E}-12(5.643 \mathrm{E}-03)$ | (-1.210E-02, 1.210E-02) | 50.00 | $-1.611 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.407 \mathrm{E}-12$ (1.190E-03) | $(-2.553 \mathrm{E}-03,2.553 \mathrm{E}-03)$ | 50.00 | $7.904 \mathrm{E}-09$ (>0.999) |
| 5 | Lower | $-9.836 \mathrm{E}-12(5.854 \mathrm{E}-03)$ | (-1.256E-02, 1.256E-02) | 50.00 | $-1.680 \mathrm{E}-09(>0.999)$ |
|  | Upper | $8.768 \mathrm{E}-12(1.373 \mathrm{E}-03)$ | (-2.946E-03, 2.946E-03) | 50.00 | $6.384 \mathrm{E}-09$ (>0.999) |
| 6 | Lower | $-9.246 \mathrm{E}-12$ (5.741E-03) | $(-1.231 \mathrm{E}-02,1.231 \mathrm{E}-02)$ | 50.00 | $-1.611 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.280 \mathrm{E}-12$ (1.271E-03) | (-2.725E-03, 2.725E-03) | 50.00 | $7.303 \mathrm{E}-09$ (>0.999) |
| 7 | Lower | $-9.500 \mathrm{E}-12(5.780 \mathrm{E}-03)$ | (-1.240E-02, 1.240E-02) | 50.00 | $-1.643 \mathrm{E}-09(>0.999)$ |
|  | Upper | $9.091 \mathrm{E}-12$ (1.296E-03) | $(-2.780 \mathrm{E}-03,2.780 \mathrm{E}-03)$ | 50.00 | $7.015 \mathrm{E}-09$ (>0.999) |
| 8 | Lower | $-9.083 \mathrm{E}-12(5.684 \mathrm{E}-03)$ | (-1.219E-02, 1.219E-02) | 50.00 | -1.598E-09 (>0.999) |
|  | Upper | $9.432 \mathrm{E}-12$ (1.217E-03) | $(-2.611 \mathrm{E}-03,2.611 \mathrm{E}-03)$ | 50.00 | $7.748 \mathrm{E}-09$ (>0.999) |
| 9 | Lower | $-9.631 \mathrm{E}-12(5.780 \mathrm{E}-03)$ | (-1.240E-02, 1.240E-02) | 50.00 | $-1.666 \mathrm{E}-09(>0.999)$ |
|  | Upper | $8.993 \mathrm{E}-12$ (1.292E-03) | $(-2.771 \mathrm{E}-03,2.771 \mathrm{E}-03)$ | 50.00 | $6.960 \mathrm{E}-09(>0.999)$ |

Appendix L
CDC Case Report Forms

## * Required Field



Hemovigilance Adverse Reaction

## Investigation Results (See Case Definition Criteria)

*Was a particular unit implicated in the adverse reaction? $\square$ YES $\square$ NO

## *Adverse reaction (Select one):

$\square$ Allergic reaction, including anaphylaxis
$\square$ Acute hemolytic transfusion reaction (AHTR):Immune Antibody: $\qquad$
$\square$ Non-immune (specify)
$\qquad$
$\square$ Delayed hemolytic transfusion reaction (DHTR):
$\qquad$
$\square$ Immune Antibody:

Non-immune (specify) $\qquad$
$\square$ Delayed serologic transfusion reaction (DSTR): Antibody: $\qquad$Febrile non-hemolytic transfusion reactionHypotensive transfusion reactionInfection A. $\square$ Bacterial (incl. sepsis)ViralOther B. Organism (specify) $\qquad$ Was a test to detect a specific pathogen performed on recipient post-transfusion? $\square$ YES $\square$ NO If YES, were any results positive or reactive? $\square$ YES Organism:_ $\square$ NO Was a test to detect a specific pathogen performed on the donor post-donation? $\square$ YES $\square$ NO If YES, were any results positive or reactive?
$\square$ YES Organism: $\qquad$ $\square$ NO Was a test to detect a specific pathogen performed on the unit post-tranfusion? (i.e., blood culture, serology, nucleic acid) $\square$ YES $\square$ NO If YES, were any results positive or reactive? $\square$ YES Organism: $\qquad$ - NO
$\square$ Post transfusion purpura (PTP)
$\square$ Transfusion associated circulatory overload (TACO)
$\square$ Transfusion associated dyspnea (TAD)
$\square$ Transfusion associated graft vs. host disease (TA-GVHD)
Has the patient received any non-irradiated blood product(s) in the past two months? $\square$ Yes $\square$ No
$\square$ Transfusion related acute lung injury (TRALI)
(Optional) Antibody studies performed:

|  | Not Done | Negative | Test result positive ( + ) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Cognate or cross reacting antigen present | No cognate or cross reacting antigen present | Not tested for cognate antigen |
| Donor or unit HLA specificity | $\square$ | $\square$ | $\square$ | $\square$ | $\square$ |
| Donor or unit HNA specificicity | $\square$ | $\square$ | $\square$ | $\square$ | $\square$ |
| Recipient HLA specificity | - | ] | $\square$ | $\square$ | $\square$ |
| Recipient HNA specificity | $\square$ | $\square$ | $\square$ | $\square$ | $\square$ |Unknown pathophysiology

$\square$ Other (specify)
i. Meets Case Definition Criteria: Def = Definitive, Pro = Probable, Pos = Possible, NA
ii. Grade: $N S=$ Non-severe, $S=$ Severe, $L T=$ Life-threatening, $D=$ Death, $N D=$ Not Determined
iii. Relationship to Transfusion: Def $=$ Definite, Pro $=$ Probable, Pos $=$ Possible, Dou=Doubtful, $R O=$ Ruled out, $N D$
*For adverse reaction selected indicate: i. Case Definition Criteria ___ ii. Grade ___ iii. Relationship

## Outcome

* $\square$ Death+ $\qquad$ Major or long-term sequelaeMinor or no sequelae $\square$ Not determined

Date of death $\qquad$ /___ 1 +Note: deaths attributable to transfusion must be reported to FDA

If recipient died, relationship of transfusion to death:DefiniteProbablePossibleDoubtfulRuled outNot determined

Appendix M
SAS Code

Available upon request

## Appendix N

Thrombocytopenia Phase III Clinical Trials

| Rank | Trial Status (Design): $\mathbf{N}$ per Treatment | Study |
| :---: | :---: | :---: |
| 1 | Completed (Parallel): 60 | Not specified |
| 2 | Completed <br> (Parallel): 240 | Safety and Efficacy of (PN-152,243)/PN-196,444 in the Prevention of Thrombocytopenia,Conditions: Non-Hodgkin Lymphoma; Hodgkin Disease; <br>  <br> Thrombocytopenia <br> Intervention: Drug: (PN-152,243)/PN-196,444 |
| 3 | Completed <br> (Parallel): 120 | $\begin{aligned} & \text { Safety and Efficacy of (PN-152,243)/PN-196,444 in the Prevention of Thrombocytopenia } \\ & \text { Condition: } \text { Sarcoma } \\ & \text { Intervention: } \text { Drug: PN-152,243)/PN- } \\ & 196,444\end{aligned}$ |
| 4 | Recruiting (1 <br> Treatment): 20 | $\begin{aligned} & \frac{\text { A Study Evaluating the Safety and Efficacy of Long-term Dosing of Romiplostim in }}{\text { Thrombocytopenic Pediatric Subjects With Immune (Idiopathic) Thrombocytopenia Purpura }} \\ & \qquad \begin{array}{ll} \text { Condition: } & \text { Thrombocytopenia in Pediatric Subjects With } \\ & \text { Immune (Idiopathic) Thrombocytopenic Purpura } \\ & \text { (ITP) } \end{array} \\ & \\ & \text { Intervention: } \\ & \text { Biological: Romiplostim } \end{aligned}$ |
| 5 | Terminated (Parallel): 292 | Eltrombopag To Reduce The Need For Platelet Transfusion In Subjects With Chronic Liver Disease And Thrombocytopenia Undergoing Elective Invasive Procedures <br> Conditions: Non-alcoholic Steatohepatitis; Chronic Liver Disease; HCV; NASH.; HIV Infection; Thrombocytopenia; Hepatitis C Virus; Hepatitis B; HIV Infections; Liver Diseases; Hepatitis B Virus Interventions: Drug: Eltrombopag; Drug: Placebo |
| 6 | Completed (1 <br> Treatment): 50 | Anticoagulant Therapy With Bivalirudin in the Performance of Percutaneous Coronary Intervention in Patients With Heparin-Induced Thrombocytopenia (AT BAT, First Inning) <br> Conditions: Heparin-Induced Thrombocytopenia; Thrombosis <br> Intervention: Drug: bivalirudin |
| 7 | $\begin{aligned} & \text { Active, not } \\ & \text { recruiting (1 } \\ & \text { Treatment): } 50 \end{aligned}$ | Angiomax in Patients With HIT/HITTS Type II Undergoing Off-Pump Coronary Artery Bypass Grafting (CABG) (CHOOSE) <br> Conditions: Thrombocytopenia; Thrombosis; Cardiac Disease; Coronary Artery Bypass Surgery <br> Intervention: Drug: Angiomax (bivalirudin) anticoagulant |
| 8 | Completed (Parallel): 63 | AMG 531 Treatment of Thrombocytopenic Subjects With Immune (Idiopathic) Thrombocytopenic Purpura (ITP) Refractory to Splenectomy <br> Conditions: Thrombocytopenia; Idiopathic Thrombocytopenic Purpura <br> Interventions: Drug: Placebo; Biological: AMG 531 |

9 Completed (Parallel): 62

10 Completed (1
Treatment): 8

Recruiting (1
Treatment): 340

2 Active, not recruiting
(Parallel): 750

3 Active, not recruiting (Parallel): 750

14 Completed (1
Treatment): 131

AMG 531 Treatment of Thrombocytopenic Subjects With Immune (Idiopathic) Thrombocytopenic Purpura (ITP) Prior to Splenectomy

| Conditions: | Thrombocytopenia; <br> Idiopathic Thrombocytopenic Purpura <br> Interventions: <br> Biological: AMG 531; Drug: Placebo |
| :---: | :--- |


| Efficacy and Safety Study of Argatroban to Treat Heparin-Induced Thrombocytopenia |  |
| ---: | :--- |
| Condition: | Heparin- |
|  | Induced Thrombocytopenia |
| Intervention: | Drug: argatroban |

Clinical Trial for Non-responders Who Previously Participated in Eltrombopag Studies TPL 103922 or TPL 108390
$\begin{array}{cl}\text { Conditions: } & \begin{array}{l}\text { Hepatitis C; } \\ \text { Thrombocytopaenia }\end{array} \\ \text { Intervention: } & \text { Drug: Eltrombopag }\end{array}$

Eltrombopag To Initiate And Maintain Interferon Antiviral Treatment To Subjects With Hepatitis C Related Liver Disease

Conditions: Hepatitis C, Chronic; Hepatitis C;
Thrombocytopenia
Intervention: Drug: eltrombopag

Eltrombopag To Initiate And Maintain Interferon Antiviral Treatment To Benefit Subjects With Hepatitis C Liver Disease

Conditions: Hepatitis C, Chronic; Hepatitis C; Thrombocytopenia
Intervention: Drug: eltrombopag

AMG 531 Versus Medical Standard of Care for Immune (Idiopathic) Thrombocytopenic Purpura
Conditions: Thrombocytopenic Purpura;
Idiopathic Thrombocytopenic Purpura;
Thrombocytopenia; Thrombocytopenia in Subjects
With Immune (Idiopathic) Thrombocytopenic
Purpura (ITP)
Interventions: Biological: AMG531;
Drug: Medical Standard of Care for ITP

Completed (1
Treatment): 50
Angiomax in Patients With HIT/HITTS Type II Undergoing CPB
Conditions: Cardiovascular Disease; Coronary Artery Bypass Surgery
Intervention: Drug: Angiomax (bivalirudin)

Recruiting (1
Treatment): 400
EXTEND (Eltrombopag Extended Dosing Study)
Conditions: Idiopathic Thrombocytopenic Purpura;
Purpura, Thrombocytopenic, Idiopathic
Intervention: Drug: eltrombopag olamine (SB-497115-GR)

17 Active, not recruiting (1 Treatment): 500

An Open Label Study of Romiplostim in Adult Thrombocytopenic Subjects With ITP
Conditions: Idiopathic Thrombocytopenic Purpura;
Thrombocytopenia; Thrombocytopenic Purpura
Intervention: Biological: Romiplostim

Completed (Parallel): 211 Treatment): 313

Active, not recruiting (1 Treatment): 40

Recruiting (Parallel): 36

Completed (1 Treatment): 60

Terminated (Parallel): 27

Completed (Parallel): 197 (Parallel): 34

Completed (Parallel): 106

Efficacy and Safety Study of Platelets Treated for Pathogen Inactivation and Stored for Up to Seven Days

Condition: Thrombocytopenia
Intervention: Device: Transfusion of Pathogen Inactivated Platelets stored for 6-7 days

Completed (1 Open Label Extension Study of AMG 531 in Thrombocytopenic Subjects With Immune (Idiopathic) Thrombocytopenic Purpura (ITP)

Conditions: Thrombocytopenia; Idiopathic Thrombocytopenic Purpura
Intervention: Biological: AMG 531

Open Label Extension Study of AMG 531 in Japanese Subjects With ITP
Condition: Thrombocytopenia in Subjects With Immune (Idiopathic) Thrombocytopenic Purpura (ITP)
Intervention: Biological: AMG 531

A Randomized Study of IVIG vs. IVIG With High Dose Methylprednisolone in Childhood ITP. Condition: Immune Thrombocytopenic Purpura
Intervention: Drug: Methylprednisolone and Intravenous Immune Globulin

Treatment of Chronic Immune Thrombocytopenic Purpura (ITP) With Intravenous Immunoglobulin IgPro10

Condition: Immune Thrombocytopenic Purpura
Intervention: Drug: Immunoglobulin Intravenous (Human )

Safety of Fondaparinux as Routine VTE Prophylaxis in Medical ICU Patients

| Condition: | Venous Thromboembolism |
| ---: | :--- |
| Interventions: | Drug: Fondaparinux; |
|  | Drug: Enoxaparin |

RAISE: Randomized Placebo-Controlled Idiopathic Thrombocytopenic Purpura (ITP) Study With Eltrombopag

| Condition: | Idiopathic Thrombocytopenic Purpur |
| ---: | :--- |
| Intervention: | Drug: eltrombopag |

P3 Study to Evaluate Efficacy and Safety of AMG 531 in Thrombocytopenic Japanese Subjects With Immune (Idiopathic) Thrombocytopenic Purpura

Condition: Idiopathic Thrombocytopenic Purpur
a
Interventions: Drug: Placebo; Drug: AMG 531

The Use of Fondaparinux in Preventing Thromboembolism in High Risk Trauma Patients
Condition: Venous Thromboembolism
Interventions: Drug: fondaparinux sodium;
Device: sequential compression devices

27 Recruiting (
Treatment): 7

IGIV Study for Chronic ITP Patients Ages 3-70
Condition: Idiopathic Thrombocytopenic Purpur a
Intervention: Drug: IGIV3I Grifols $10 \%$

Complementary Treatment of PG2 to Improve Clinical Benefit Response and Quality of Life in Fatigue

Conditions: Quality of Life; Fatigue; Complementary Intervention: Drug: PG2

Completed Efficacy and Safety of Low-molecular Weight Heparin for Thromboprophylaxis in Acutely Ill (Parallel): 342

Medical Patients

| Condition: | Embolism |
| ---: | :--- |
| Interventions: | Drug: Certoparin; |
|  | Drug: Heparin |

Myfortic vs. Cellcept in Kidney Transplant Recipients
Completed
(Parallel): 150

(Parallel): 120

## Completed

 (Parallel): 1329Recruiting (Parallel): 500

Karenitecin Versus Topotecan in Patients With Advanced Epithelial Ovarian Cancer Condition: Ovarian Cancer Interventions: Drug: Karenitecin; Drug: Topotecan

Study Comparing Desirudin With Heparin to Prevent Vein Clots After Heart and Lung Surgery Condition: Deep Venous Thrombosis Intervention: Drug: Desirudin (Iprivask ${ }^{\text {TM }}$ )

First Line IRESSA ${ }^{\text {TM }}$ Versus Carboplatin/Paclitaxel in Asia
Condition: Non-small Cell Lung Cancer Interventions: Drug: Gefitinib; Drug: Carboplatin; Drug: Paclitaxel
Conditions: Quality of Life; Fatigue; Complementary

Condition: Embolism
Drug: Heparin

Condition: End Stage Renal Disease
Interventions: Drug: Mycophenolate Sodium Delayed Release Tablets; Drug: Mycophenolate Mofetil

## Coronary Artery Bypass Graft Surgery: The Fonda CABG Study

Condition: Coronary Bypass Graft Failure/Occlusio n
Intervention: Drug: Fondaparinux

Terminated (1 Safety and Efficacy Study To Compare Uniplas With Cryosupernatant Plasma In Thrombotic Treatment): $112 \quad$ Thrombocytopenic Purpura (TTP)

Condition: Thrombotic Thrombocytopenic Purpura (TTP )
Intervention: Drug: Uniplas

37 Recruiting (1
Treatment): 20

38 Completed (1
Treatment): 19

39 Active, not
recruiting (1
Treatment): 19

40 Completed
(Parallel): 103

41 Completed (Paired

Clinical Evaluation of SB-497115-GR in Chronic Idiopathic Thrombocytopenic Purpura (ITP)

| Condition: | Chronic Idiopathic Thrombocytopenic Purpura |
| ---: | :--- |
| Interventions: | Drug: SB-497115-GR 12.5mg tablet; Drug: SB- |
|  | $497115-$ GR 25mg tablet; Drug: SB-497115- |
|  | GR 12.5 mg matching placebo tablet |


| Clinical Trial in Patients Diagnosed With Immune Thrombocytopenic Purpura |  |  |
| :--- | :---: | :---: |
| Condition: |  | Idiopathic Thrombocytopenic Purpur |
|  |  |  |
|  |  |  |
| a |  |  |
| Intervention: |  |  |
| Biological: IGIV3I Grifols |  |  |

Ig NextGen 10\% in Idiopathic Thrombocytopenic Purpura (ITP) Patients
Condition: Idiopathic Thrombocytopenic Purpura (ITP )
Intervention: Drug: IgNextGen 10\%

| Clinical Evaluation of Eltrombopag in Chronic Idiopathic Thrombocytopenic Purpura (ITP) |  |
| ---: | :--- |
| Conditions: | Idiopathic Thrombocytopenic Purpura; |
|  | Purpura, Thrombocytopenic, Idiopathic |
| Intervention: | Drug: Eltrombopag oral tablets |

A Study Evaluating the Addition of MabThera (Rituximab) to Standard Treatment in Patients With Idiopathic Thrombocytopenic Purpura (ITP)

Condition: Idiopathic Thrombocytopenic Purpur
a
Intervention: Drug: rituximab

Crossover): 100

42 Completed (1 Treatment): 116

3 Completed (Paired Crossover): NA

Safety and Efficacy Study of a 10\% Intravenous Immune Globulin Solution in Subjects With Primary Immunodeficiency Disorders

Conditions: Primary Immunodeficiency Disorders; Immune Thrombocytopenic Purpura (ITP); Kawasaki Syndrome
Intervention: Procedure: Immune Globulin Intravenous (Human), 10\%

Study of Higher Dose of Rituxan Versus Standard Doses of Rituxan With Cyclophosphamide, Vincristine, and Prednisone in Subjects With Chronic ITP

Condition: Immune Thrombocytopenic Purpura
Interventions: Drug: Rituxan and Cyclophosphamide, Vincristine
and Prednisone; Drug: Higher Dose of Rituximab

45 Terminated
(Parallel): 2
Evaluating the Effectiveness of Adding Rituximab to Standard Treatment for Thrombotic Thrombocytopenic Purpura (TTP) (The STAR Study)

Condition: Thrombotic Thrombocytopenic Purpura Interventions: Drug: Rituximab; Procedure: Plasma exchange;

Drug: Corticosteroids


[^0]:    ${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas; ${ }^{4}$ Program 32.sas

[^1]:    ${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas; ${ }^{3}$ Program 31.sas; ${ }^{4}$ Program 32.sas

[^2]:    ${ }^{1}$ Shapiro-Wilk test for normality of the effect; ${ }^{2}$ Program 30.sas

