

Medical University of South Carolina

MEDICA

MUSC Theses and Dissertations

2018

Missing Data Methods for ICU SOFA Scores in Electronic Health Records Studies: Results from a Monte Carlo Simulation Study

Daniel Lee Brinton

Medical University of South Carolina

Follow this and additional works at: <https://medica-musc.researchcommons.org/theses>

Recommended Citation

Brinton, Daniel Lee, "Missing Data Methods for ICU SOFA Scores in Electronic Health Records Studies: Results from a Monte Carlo Simulation Study" (2018). *MUSC Theses and Dissertations*. 311.
<https://medica-musc.researchcommons.org/theses/311>

This Dissertation is brought to you for free and open access by MEDICA. It has been accepted for inclusion in MUSC Theses and Dissertations by an authorized administrator of MEDICA. For more information, please contact medica@muscd.edu.

MISSING DATA METHODS FOR ICU SOFA SCORES IN ELECTRONIC HEALTH
RECORDS STUDIES: RESULTS FROM A MONTE CARLO SIMULATION STUDY

BY

Daniel Lee Brinton

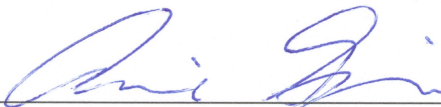
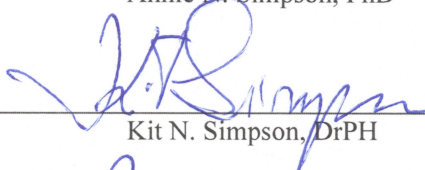
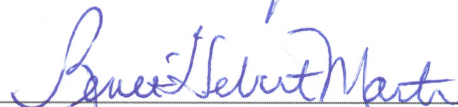
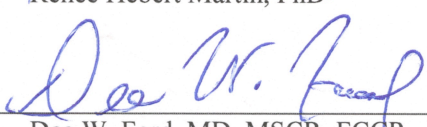
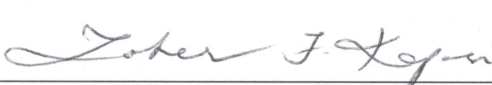
A dissertation submitted to the faculty of the Medical University of South Carolina
in partial fulfillment of the requirements for the degree
Doctor of Philosophy
in the College of Health Professions

MISSING DATA METHODS FOR ICU SOFA SCORES IN ELECTRONIC HEALTH RECORDS STUDIES: RESULTS FROM A MONTE CARLO SIMULATION STUDY

BY

Daniel Lee Brinton

Approved by:

Chair, Project Committee	 Annie N. Simpson, PhD	7/13/18 Date
Member, Project Committee	 Kit N. Simpson, DrPH	7/13/18 Date
Member, Project Committee	 Renée Hebert Martin, PhD	7/13/18 Date
Member, Project Committee	 Dee W. Ford, MD, MSCR, FCCP	7/13/18 Date
Dean, College of Health Professions	 Zoher F. Kapasi, PhD, PT, MBA	07/23/2018 Date

Acknowledgements

I thank my parents, Douglas and Jackie Brinton, for raising me and my sister to value education and encouraging us to pursue our dreams. Your years of investment in Angela and me, ensuring we had exposure to many of the finest things in life—the outdoors, music, the arts—laid a great foundation upon which to build.

Dr. Doug Henry, my mentor from middle school through high school, because of your dedication to helping others and the selfless investment in teaching and coaching others you give tirelessly of yourself to help others realize their full potential. The latent ember of intellectual curiosity that existed within me you saw and stoked through your mentorship in computer programming, exposure to some of the amazing products of computer programming such as the chemical modeling software you helped develop at MDL, and sage career advice. I am forever indebted to you.

I thank Dr. Ralph Ward for his assistance throughout the program, helping me to access VA data for other projects, and helping me to troubleshoot the algorithms for missing data generation. I thank Dr. Wenle Zhao for his methodological insight into simulation studies.

I thank my mentor, whom I first met as the Professor for our Comparative Effectiveness Research course in the MHA program, Dr. Annie Simpson. Your passion for research, dedication, and affinity for teaching were inspirational from the start. When my grandfather was diagnosed with lung cancer during the CER course in 2013 I desperately searched the literature to help understand the gravity of that diagnosis and how much more time I might have with him. You helped me to interpret the literature so that I understood what the future might hold for my grandfather. This galvanized my desire to someday enter a PhD program to prepare to conduct health services research. While I may not be a clinician and able to directly help patients, becoming a health services researcher, I knew, would allow me to help build the body of knowledge that informs good clinical practice and makes a lasting difference to patients. Thank you for believing in me and choosing me to be your first PhD mentee.

To the members of my committee, Drs. Kit Simpson, Renée Martin, and Dee Ford, thank you for your mentorship, sharing of your expertise, and professionalism. Our meetings were always exciting, synergistic, and something from which I walked away with renewed dedication to my research. I look forward to our continued research together.

I thank my wife Melissa Brinton for her support, encouragement, and flexibility in scheduling our lives over the past few years. Although you, yourself are also a full-time graduate student, you lifted the load—caring for our son Judah when I needed to be away for research, conferences, and the writing of this dissertation; you are the quintessential spouse and my best friend. I am further indebted to you for your clinical insight, offering ad hoc consultations and helping me to stay grounded in this research by constantly reminding me to keep focus on the patients for whom we conduct research. This work is dedicated to you.

Finally, my son Judah. When I started this journey you were but a newborn and now you are a caring, jovial, energetic, and delightful boy. You have provided the humor and encouragement I needed during this time.

Abstract of Dissertation Presented to the
Doctor of Philosophy Program in Health and Rehabilitation Science
Medical University of South Carolina
In Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

MISSING DATA METHODS FOR ICU SOFA SCORES IN ELECTRONIC HEALTH
RECORDS STUDIES: RESULTS FROM A MONTE CARLO SIMULATION STUDY

by

Daniel Lee Brinton

Chairperson: Annie N. Simpson, PhD
Committee: Kit N. Simpson, DrPH
Renée Hebert Martin, PhD
Dee W. Ford, MD, MSCR, FCCP

This study utilizes electronic health record data from the Medical University of South Carolina's intensive care units as the basis for this Monte Carlo simulation study—which compares four methods for handling missing SOFA scores, both at the composite and component levels. The four methods examined herein include: complete case analysis, median imputation, zero imputation (the method recommended by the creators of the SOFA score), and multiple imputation. This study found that zero imputation introduced the most bias across all three outcomes studied, and therefore is not recommended. Complete case analysis, or ignoring missing data, caused varying amounts of bias—as did median imputation. Multiple imputation, on the other hand, performed well for all three outcomes studied, both at the composite and component levels, demonstrating this method's superior value in the presence of missing SOFA scores.

TABLE OF CONTENTS

Acknowledgements.....	ii
List of Acronyms	vii
List of Figures.....	viii
List of Tables	xvi
1 INTRODUCTION	1
2 REVIEW OF THE LITERATURE.....	5
2.1 Problem Statement: Why Missing Data are Problematic.....	5
2.2 Current State of Handling Missing Data in Health Services Research.....	7
2.3 Electronic Health Record Data for Research	10
2.4 Clinical Background – Respiratory Failure	13
2.4.1 Respiratory Failure: Population Statistics.....	13
2.4.2 Ventilator-Dependent Respiratory Failure.....	14
2.4.3 Respiratory Failure: Pathology	15
2.5 Severity of Illness Scoring.....	15
2.5.1 SOFA Score	16
2.5.2 Interpretation of SOFA Score	21
2.5.3 Examples of SOFA Score in the Literature.....	23
2.5.4 Missingness of SOFA Score Items	23
2.6 Mechanisms of Missingness	24
2.6.1 Missing Completely at Random (MCAR)	25
2.6.2 Missing at Random (MAR).....	25
2.6.3 Missing Not at Random (MNAR).....	26
2.6.4 Tests for Missing Data Mechanism	27
2.7 Missing Data Patterns	28
2.8 Amount of Missing Data.....	29
2.9 Analytical Approaches to Missing Data	29
2.9.1 Adjustment Methods.....	30
2.9.2 Imputation Methods.....	32
2.9.2.1 Deterministic Imputation	32
2.9.2.2 Single Imputation.....	34
2.9.2.3 Multiple Imputation	35
2.9.3 Likelihood Methods.....	39

2.9.3.1	Maximum Likelihood Estimation with EM Algorithm	40
2.9.3.2	Full Information Maximum Likelihood	41
2.9.4	Methods Available in Common Statistical Software	41
2.10	Recapitulation	44
3	METHODS	45
3.1	Specific Aims and Hypotheses	45
3.2	Data Source	46
3.3	Study Population	47
3.4	Statistical Software and Data Management	48
3.5	Methods for Multiple Item Instruments	48
3.6	Aim 1 – Univariate Missingness (SOFA Score, Composite Level).....	50
3.7	Aim 2 – Multivariate Missingness (SOFA Score, Item Level).....	50
3.8	Simulation Process & Outcomes Analysis.....	51
3.8.1	Simulation Algorithm	51
3.8.2	Simulation Parameters	52
3.8.2.1	Missing Data Mechanism & Generation of Missing Data	53
3.8.2.2	Assignment of Missing Data Patterns	55
3.8.2.3	Simulation Runs & Percent Missingness	55
3.8.3	Missing Data Methods	56
3.8.3.1	Method 1: Complete Case Analysis.....	57
3.8.3.2	Method 2: Median Imputation	58
3.8.3.3	Method 3: Imputation per SOFA Guidelines (Zero).....	58
3.8.3.4	Method 4: Multiple Imputation.....	59
3.8.4	Analysis of Outcomes	61
3.8.4.1	Outcome 1: Death	62
3.8.4.2	Outcome 2: Total Charges	63
3.8.4.3	Outcome 3: ICU Length of Stay	65
3.8.5	Output of Results from Simulations.....	66
3.8.6	Assessment of Simulation	67
3.8.7	Monitoring of Simulation Process	68
4	RESULTS	71
4.1	Data Used in Dissertation	71
4.1.1	Descriptive Characteristics	73
4.1.2	Bivariate Analyses	75
4.1.3	SOFA Scores.....	76
4.1.3.1	Missing Data Mechanism.....	76

4.1.3.2	Missing Data Patterns	79
4.1.3.3	Distribution of SOFA Scores	82
4.2	Fully-Observed Dataset Outcomes	83
4.2.1	Outcome 1: Death	84
4.2.2	Outcome 2: Total Charges	87
4.2.3	Outcome 3: ICU Length of Stay	91
4.3	Aim 1 – Results.....	94
4.3.1	Outcome 1: Death	95
4.3.2	Outcome 2: Total Charges	102
4.3.3	Outcome 3: ICU Length of Stay	109
4.4	Aim 2 – Results.....	116
4.4.1	Outcome 1: Death	116
4.4.2	Outcome 2: Total Charges	123
4.4.3	Outcome 3: ICU Length of Stay	130
4.5	Summary and Comparison of Results.....	137
5	DISCUSSION.....	139
5.1	Integration of Findings.....	139
5.2	Limitations	142
5.3	Future Research	143
APPENDICES		145
Appendix A. Analytical approaches to missing data, search terms		146
Appendix B. Performance of missing data methods, tables.....		147
Appendix C. Correlation table		153
Appendix D. Example SAS Code.....		155
REFERENCES		162

List of Acronyms

ACA	Available case analysis
APACHE	Acute physiology age and chronic health evaluation system
AUROC	Area under the receiver operating characteristic curve
BMI	Body mass index
CAM-ICU	Confusion assessment method for the ICU
CCA	Complete case analysis
CDW	Clinical data warehouse
EB	Entropy balance
EM	Expectation maximization algorithm for maximum likelihood estimation
EHR	Electronic health record
FCS	Fully-conditional specification (another term for MICE)
GCS	Glasgow coma scale
GDP	Gross domestic product
FIML	Full information maximum likelihood
HCUP	Healthcare cost and utilization project
HSR	Health services research
ICD-9-CM	International Classification of Diseases, 9 th revision, clinical modification
ICD-10-CM	International Classification of Diseases, 10 th revision, clinical modification
ICU	Intensive care unit
IPW	Inverse probability weighting
NIS	Nationwide Inpatient Sample
LOCF	Last observation carried forward
MAP	Mean arterial pressure
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov-chain Monte Carlo multiple imputation (an implementation of MVNI)
MI	Multiple imputation
MICE	Multivariate imputation by chained equations
MIM	Missing indicator method
MLE	Maximum likelihood estimation
MNAR	Missing not at random
MODS	Multiple organ dysfunction syndrome
MVNI	Multivariate normal imputation
PICU	Pediatric intensive care unit
RASS	Richmond Agitation-Sedation Scale
SBT	Spontaneous breathing trial
SIRS	Systemic inflammatory response syndrome criteria
SOFA	Sequential organ failure assessment
VDRF	Ventilator-dependent respiratory failure

List of Figures

Figure 2.1 Maximum SOFA score vs. in-ICU mortality rate [84].....	22
Figure 2.2 Complete data matrix	24
Figure 2.3 Missing data patterns, monotonic vs. non-monotonic	28
Figure 2.4 Taxonomy of analytical approaches to missing data.....	30
Figure 2.5 Conceptual diagram of multiple imputation.....	37
Figure 3.1 Simulation algorithm	52
Figure 4.1 Data flow diagram	72
Figure 4.2 Histogram of SOFA scores in the fully-observed dataset	82
Figure 4.3 Forest plot of the odds ratios and 95% confidence intervals for predicting in-hospital death in the fully-observed dataset	86
Figure 4.4 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the logistic regression model predicting <i>Death</i> (Aim 1 – Composite Level).....	97
Figure 4.5 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MAR</i> missing data mechanism (Aim 1 – Composite Level).....	98
Figure 4.6 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 1 – Composite Level)	98
Figure 4.7 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 1 – Composite Level)	99
Figure 4.8 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 1 – Composite Level)...	99

Figure 4.9 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MAR</i> missing data mechanism (Aim 1 – Composite Level)	100
Figure 4.10 Comparison of efficiency estimates for the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 1 – Composite Level)	100
Figure 4.11 Comparison of efficiency estimates for the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 1 – Composite Level)	101
Figure 4.12 Comparison of efficiency estimates for the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 1 – Composite Level).....	101
Figure 4.13 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> (Aim 1 – Composite Level)	104
Figure 4.14 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MAR</i> missing data mechanism (Aim 1 – Composite Level).....	105
Figure 4.15 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 1 – Composite Level)	105
Figure 4.16 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 1 – Composite Level).....	106

Figure 4.17 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 1 – Composite Level)	106
Figure 4.18 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MAR</i> missing data mechanism (Aim 1 – Composite Level)	107
Figure 4.19 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 1 – Composite Level)	107
Figure 4.20 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 1 – Composite Level)	108
Figure 4.21 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 1 – Composite Level).....	108
Figure 4.22 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> (Aim 1 – Composite Level)	111
Figure 4.23 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MAR</i> missing data mechanism (Aim 1 – Composite Level)	112
Figure 4.24 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 1 – Composite Level).....	112

Figure 4.25 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 1 – Composite Level).....	113
Figure 4.26 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 1 – Composite Level).....	113
Figure 4.27 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MAR</i> missing data mechanism (Aim 1 – Composite Level).....	114
Figure 4.28 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 1 – Composite Level).....	114
Figure 4.29 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 1 – Composite Level).....	115
Figure 4.30 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 1 – Composite Level).....	115
Figure 4.31 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the logistic regression model predicting <i>Death</i> (Aim 2 – Component Level)	118
Figure 4.32 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting <i>Death</i> , with the <i>MAR</i> missing data mechanism (Aim 2 – Component Level).....	119

Figure 4.33 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level).. 119

Figure 4.34 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level) 120

Figure 4.35 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level) 120

Figure 4.36 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MAR* missing data mechanism (Aim 2 – Component Level)..... 121

Figure 4.37 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level).. 121

Figure 4.38 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level) 122

Figure 4.39 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level) 122

Figure 4.40 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges* (Aim 2 – Component Level)..... 125

Figure 4.41 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MAR</i> missing data mechanism (Aim 2 – Component Level)	126
Figure 4.42 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 2 – Component Level).....	126
Figure 4.43 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 2 – Component Level).....	127
Figure 4.44 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 2 – Component Level).....	127
Figure 4.45 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MAR</i> missing data mechanism (Aim 2 – Component Level).....	128
Figure 4.46 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 2 – Component Level).....	128
Figure 4.47 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 2 – Component Level).....	129
Figure 4.48 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting <i>Total Charges</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 2 – Component Level)	129

Figure 4.49 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> (Aim 2 – Component Level).....	132
Figure 4.50 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MAR</i> missing data mechanism (Aim 2 – Component Level).....	133
Figure 4.51 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 2 – Component Level)	133
Figure 4.52 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 2 – Component Level)	134
Figure 4.53 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Right</i> missing data mechanism (Aim 2 – Component Level)	134
Figure 4.54 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MAR</i> missing data mechanism (Aim 2 – Component Level)	135
Figure 4.55 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Left</i> missing data mechanism (Aim 2 – Component Level).....	135
Figure 4.56 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting <i>ICU Length of Stay</i> , with the <i>MNAR Middle</i> missing data mechanism (Aim 2 – Component Level).....	136

Figure 4.57 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)..... 136

List of Tables

Table 1 SOFA score calculation	18
Table 2 Scoring of the Glasgow Coma Scale (GCS).....	20
Table 3 Missing data methods available in SAS, SPSS, and Stata	43
Table 4 ICD-9 and ICD-10 procedure codes for study inclusion	48
Table 5 Example calculations of the SOFA score	49
Table 6 Simulation output table (example).....	67
Table 7 Demographics and characteristics of patients in the original dataset	74
Table 8 Interpretation of Spearman's rank correlation coefficients.....	76
Table 9 Odds ratios and 95% confidence intervals for predicting a SOFA score being missing in the original data	78
Table 10 Frequency of missing SOFA score components in original data.....	80
Table 11 Twenty-five most common missing data patterns	81
Table 12 Distribution of SOFA scores in the fully-observed dataset	83
Table 13 Odds ratios and 95% confidence intervals for predicting in-hospital death in the fully-observed dataset.....	85
Table 14 Differences (expressed as a ratio) between the point estimates and 95% confidence intervals for Total Charges in comparison to reference groups in the fully-observed dataset.....	89
Table 15 Least squares means exponentiated point estimates and 95% confidence intervals for Total Charges, expressed in thousands of dollars, in the fully-observed dataset.....	90
Table 16 Differences (expressed as a ratio) between the point estimates and 95% confidence intervals for ICU Length of Stay in comparison to reference groups in the fully-observed dataset	92
Table 17 Least squares means exponentiated point estimates and 95% confidence intervals for ICU Length of Stay, expressed in days, in the fully-observed dataset	93
Table 18 Coverage of the 95% confidence interval for various missing data methods (Aim 1) .	147

Table 19 Coverage of the 95% confidence interval for various missing data methods (Aim 2) .	148
Table 20 Relative bias for various missing data methods (Aim 1)	149
Table 21 Relative bias for various missing data methods (Aim 2)	150
Table 22 Efficiency for various missing data methods (Aim 1)	151
Table 23 Efficiency for various missing data methods (Aim 2)	152
Table 24 Intercorrelations for variables used in this study, measured by Spearman’s rank correlation coefficient, ρ_s	154

1 INTRODUCTION

The cost of caring for critically-ill patients has grown from \$55.5 billion in 2000 [1] to \$81.7 billion in 2005 [2]. The increase in expenditures on critical care medicine from 2000-2005 represents an 17.9% increase in percent of gross domestic product (GDP) over a five-year span, from 0.56% to 0.66% of GDP, accounting for 4.1% of health expenditures nationwide—demonstrating the costs are growing in comparison with overall national expenditures. This rapid increase in expenditures on critical care could be partially attributed to the rise in incidence of mechanical ventilation amongst adults in intensive care units (ICU), which increased from 284 per 100,000 adults in 1996 to 314 per 100,000 in 2002 [3]. While these figures are in need of refreshing with more recent estimates, they illustrate the magnitude of money that is spent in one area of medicine, critical care medicine, and how this area has a measurable impact on our nation's budget.

Admission to an ICU not only has large financial consequences, sequelae of stays in the ICU also manifest. One such outcome is the development of acute post-traumatic stress disorder (PTSD) related symptoms, possibly due to delirium during the patient's ICU stay [4, 5]. Approximately 1 in every 5 ICU survivors have clinically-significant PTSD symptoms within 12 months of ICU discharge [6]. Several recommendations for decreasing the likelihood of PTSD symptoms and other psychiatric morbidities exist. The first recommendation includes offering lighter amounts of sedation to improve patient recall [7]. Another recommendation is to have ICU diaries, written in the second person in patient-friendly language by clinicians caring for the patient and family members [6]. Finally, another recommendation is to use the ABCDEF bundle (Assess, prevent, and manage pain; Both spontaneous awakening trials and spontaneous breathing trials; Choice of analgesia and sedation; Delirium: assess, prevent, and manage; Early mobility and exercise; and Family engagement and empowerment [8]) to improve outcomes [5]. Other

potential sequelae of ICU stays include anxiety, depression, cognitive impairments, family and social network distress, sleep abnormalities, general distress, and diminished quality of life [9].

The problems of increasing costs of caring for the critically ill, as well as comorbidities associated with that care, drive the need for research to improve patient outcomes and reduce overall costs. This is accomplished through both interventional and retrospective studies. Interventional studies within the intensive care unit are increasingly using designs such as the pragmatic cluster-randomized stepped wedge design. This design specifies that all clusters—in this case ICUs—will be randomly crossed over from the control group to the intervention group [10]. In both interventional and retrospective outcomes studies, use of a patient severity score—such as the Sequential Organ Failure Assessment (SOFA) score—is vital in multivariable models to control for baseline patient severity. While caution has been given for using these patient severity scores on the individual level for prognosis, they work well for severity adjustment and case-mix adjustment [11]. Therefore, the use of severity score systems for ICU patients is common, but not without imperfections in their execution.

Severity scores such as the SOFA score are component scores of multiple datapoints. In the case of the SOFA score, there are 6 items that are physiological clues of organ failure, such as platelet counts (indicative of coagulation dysfunction), bilirubin levels (indicative of liver dysfunction), and the Glasgow Coma Scale (indicative of central nervous system dysfunction). It is not uncommon that one value may be missing from the medical record, preventing the calculation of the SOFA score.

When this baseline measurement of disease severity is missing in retrospective observational studies, patients may be excluded from the analysis as either an *a priori* methodological decision or inadvertently through complete case analysis. The result of this methodological choice has the potential to bias the study's results and will certainly decrease the statistical power to find a

difference in groups, should a difference exist. As the utilization of EHR data is necessarily retrospective—there is nothing that a researcher can do to improve the rates of data collection—focus needs to be given to methods of dealing with these missing data, rather than preventing missing data.

This study has been designed to examine the effects of missing SOFA score data in retrospective observational studies that use electronic health record data capturing patient stays in the intensive care unit for ventilator-dependent respiratory failure (VDRF) to accomplish the following:

- 1) Ascertain the degree to which results may be biased at various percentages of missing SOFA score data
- 2) Examine methods of dealing with missing data that are commonly available in statistical software packages used by Health Services Researchers

AIM 1

To examine the impact of missing SOFA score data on ICU clinical outcomes studies among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data at the composite score level.

AIM 2

To examine the impact of missing SOFA score data on ICU clinical outcomes studies among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data at the component item level.

Rationale of Importance (AIMS 1 & 2)

Examination of techniques for handling missing SOFA score data that are available in statistical software that is most commonly used by applied researchers—SAS, SPSS, and Stata—will provide valuable insight to researchers and clinicians. The insights this study will provide include suggestions for the best methods of handling missing data; whether missing data should be handled at the component level of the SOFA score or at the composite level of the SOFA score; and how missingness of the SOFA score affects various outcomes. Finally, the findings of these aims will provide researchers with guidelines for determining at what percentage of missingness of the SOFA score should one be worried, as currently most recommendations for missing data are ballpark figures and are not specific to the SOFA score in particular, nor ICU severity scoring systems in general.

2 REVIEW OF THE LITERATURE

2.1 Problem Statement: Why Missing Data are Problematic

Missing data present a special problem in statistics in that it is impossible to use numbers that are not present. While this may seem obvious, its implications are serious as data which are missing have the potential to seriously bias results—possibly resulting in inaccurate estimates of effect size and even direction. While the general advice is to avoid missing data at all costs, this point is irrelevant in areas where research uses secondary data sources—such as billing or electronic health records (EHR). For researchers who use secondary data sources, the only option is to deal as effectively as one can with the available data to try to answer research questions with as much accuracy and precision as possible.

In multivariable regression models, where multiple independent variables (or covariates) are contributing toward explaining a single dependent variable (or outcome) the problem with missing data may not even be realized by the researcher. If one covariate's value is missing in a regression analysis the default behavior in all major statistical software is to simply exclude this entire observation from the analysis. This exclusion is indifferent to the importance of the missing value—that covariate may offer little (or very much) explanatory value to the model. The exclusion is also indifferent to the amount of other data available in that observation; there may be scores of other covariates that provide rich information toward explaining the outcome variable. Yet the entire observation is excluded from the analysis due to a missing value in one covariate—something termed complete case analysis (CCA). While missingness of less than 5% is considered trivial [12], the amount of missing data and the implications of omitting these observations from analysis needs attention.

In other cases, the problem with missing data may be realized by the researcher—yet ignored due to the large number of cases available for analysis. This is the case with research

that uses electronic medical records of large health systems or research that uses billing records. In these cases, sufficient numbers of cases with complete covariate data may exist for analysis—resulting in a study that is well powered to find a difference if one exists, even with excluding lots of records due to missing data. The researcher simply makes the decision to set a study inclusion criterion that all cases must have complete data. Little consideration to the amount and types of missing data is given in presentation of how the cohort was developed. It is no wonder Paul Allison (2009) describes missing data as the, “dirty little secret of statistics” (p. 72).

Reporting guidelines are available to aid HSR studies to improve the quality and aid reproducibility of research. One guideline is the Strengthening The Reporting of Observational Studies in Epidemiology (STROBE), whose purpose is offering full disclosure in the analysis to allow for reproducibility and provide candor. STROBE underscores the importance of handling missing data—requiring an explanation of how missing data were handled along with an explanation of the number of subjects with missing data for each item of interest with the methods section of a peer-reviewed paper [13]. Another guideline is the Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD), an extension of STROBE, upholds all the requirements of STROBE but adds the requirement of discussing the implications of missing data in the limitations section of the paper [14]. In fact, the International Conference on Medical Journal Editors (ICMJE) highly encourages that journal articles be submitted with completed guideline checklists, such as those discussed above [15]. Unfortunately, the International Society for Pharmacoeconomics and Outcomes Research’s (ISPOR) guidelines for retrospective database studies merely mentions missing data as a quality check of the data source, and not as a potential source of bias [16]. However, a recent taskforce report from ISPOR and the International Society for Pharmacoepidemiology (ISPE)

acknowledges that missing data are a threat to validity that should be addressed [17], and offers suggestions for how describing how missing data were handled [18]. Finally, the Patient-Centered Outcomes Research Institute (PCORI) published methodology standards for scientifically valid patient-centered outcomes research [19]. There are 12 detailed standards, with one entire standard devoted toward the prevention and handling of missing data.

While most of the guidelines provide a cautious set of recommendations for handling missing data, all of them mention it as an item for methodological and statistical consideration—demonstrating the importance of properly handling missing data in health services research studies.

2.2 Current State of Handling Missing Data in Health Services Research

Unfortunately, methods for handling missing data are not widely used in health services research. One review of the utilization of multiple imputation (MI) in two top-tier journals—The Lancet and New England Journal of Medicine—over a 6-year period (2008 through 2013) found only 103 articles that used MI, 45 in NEJM and 58 in The Lancet [20]. Of these 103 studies only 30 (29.1%) were observational with 11 (10.7%) being studies using routinely collected data. The study also found nearly all the papers handled these data with insufficient rigor. The study does not give the total number of studies in the two journals during that timeframe that were evaluated, which would have informed readers about the incidence of MI during the time examined. However, a manual search for this dissertation revealed 1,373 research articles in NEJM and 1,064 in The Lancet, totaling 2,437 articles. This shows that only 4.2% of articles in these two top-tier journals used MI; 3.3% in NEJM, 5.5% in The Lancet. This suggests that missing data are a somewhat rare phenomenon, perhaps being dealt with methodologically in the analysis (but not reported), or are simply being ignored; the latter of these possibilities, rather than the former is more likely.

Therefore, a more systematic approach to help understand how well health services researchers are doing with handling missing data is warranted. In the next section, an examination of missing data approaches will be conducted to determine the extent to which missing data techniques are utilized, or if missing data are mentioned anywhere in the paper, for a one-month period of all literature that used a popular commercial claims database.

Truven Health Analytics MarketScan® research databases are widely-used for health services research. The most commonly used MarketScan databases are the Commercial and Medicare supplemental databases. These databases provide de-identified health insurance claims across the continuum of care (e.g. inpatient, outpatient, outpatient pharmacy, carve-out behavioral healthcare) as well as enrollment data from roughly 350 large employers and health plans across the United States who provide private healthcare coverage for more than 50 million employees, their spouses, dependents, and Medicare-eligible retirees with supplemental plans [21]. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans. In total, there are more than 20 billion service records in these databases spanning back to 1995. A review of research published using MarketScan databases provides illustration of the breadth, depth, and quality of research being conducted in health services research.

In August 2017, a review of all papers published in January of 2017 that used MarketScan research databases was conducted to determine the extent to which missing data techniques were utilized, or if missing data were mentioned anywhere in the paper. To locate these papers a search of the Ovid MEDLINE database for the terms *Truven* or *MarketScan* in the title or abstract was conducted, which yielded 18 papers. These papers were then searched for any of the terms listed in Appendix A—which includes terms for all the major analytical approaches to handling missing data, as well as indicators that missing data were considered (e.g. the term

missing). These papers were also manually reviewed to ascertain if missing data were mentioned, or if addressed using any analytical technique for handling missing data.

Of these 18 papers, 13 (72.2%) did not mention missing data or any analytical approaches to handling missing data anywhere in the text [22-34]; only 5 papers (27.8%) mentioned missing data at all [35-39]. Of these 5 papers, 4 collapsed categories in potential covariates to include missing with another category (e.g. *Other/Missing*) [35, 36, 38, 39] and one used missingness as an exclusion criterion without discussion of potential bias as a result of that decision [37]. In summary, none of the papers investigated over this one-month sample of time utilized a satisfactory method of handling missing data. This illustrates the common practice within health services research when using large secondary data sources—such as billing data—of simply excluding subjects with missing data, or when the data are categorical, of lumping together missing with the smaller categorical groups.

In summary, the case has been made that missing data are problematic for research studies as they can bias the findings. An examination into the state of handling missing data in health services research has also been made, showing a large area for improvement in using missing data methods. In sections that follow we will examine the clinical syndrome through which we explore missing data, examining respiratory failure and a severity of illness scoring system used for intensive care unit patients. We will then explore electronic health record data as a data source for observational research. Finally, we will take a look at the statistical topic of handling missing data, to include mechanisms for missingness, missing data patterns, amount of missing data, analytical approaches available to deal with missing data, and finally existing guidelines for working with missing data.

2.3 Electronic Health Record Data for Research

Electronic health record (EHR) data as a source for research purposes was promised to be the “golden egg” for research. A seemingly unlimited amount of data available to researchers for asking clinical research questions held the promise of forever changing research. The thought was that since the full record was available to researchers for retrospective studies, this would allow any pertinent question to be asked as all the key information—diagnoses, medications administered, radiological tests, and laboratory results—are available. This opened the door for a variety of research, including comparative effectiveness research, rare disease research, and evaluation of quality improvement initiatives.

Prior to the availability of EHR data for research, retrospective studies relied on previously recorded data, such as billing records or clinical trials data that was being repurposed to pose new research questions. These data sources have their limitations, as they were created for a specific purpose and are subject to their inherent limitations. However, EHR data are different as the entire clinical picture of care for populations of patients is seemingly captured; one would think that everything that is pertinent is available within the EHR—just waiting to be queried. Combining quasi-experimental techniques that minimize selection bias, such as propensity score matching [40, 41], with this rich source of EHR data further underscored the possibility of conducting causal analyses using these retrospective data. Further, having EHRs would transform healthcare by, in part, allowing implementation of research findings for disease prevention and chronic disease management [42]. The future looked bright for research, but reality had not yet set in.

Unfortunately, electronic health record data are a bit more difficult to work with than administrative data, due to the nature of these data. Administrative data are already coded with diagnosis and procedure codes. However, EHR data—while also having this information—has

much more information that is unstructured in free-text fields, such as provider notes, as well as bedside and laboratory data that are harder to turn into discrete fields by which research questions can be asked. For free-text fields, natural language processing is required for automated chart review. This approach requires validation for each disease or condition being phenotyped, and is subject to common problems—such as misspellings, abbreviations, and negation terms (e.g., “absence of hepatocellular nodules”) that might otherwise give a false-positive [43].

A systematic literature review of all health outcomes research studies conducted in the United States from 2000-2007 that used EHR data was conducted in 2009 to examine how EHRs were being used for outcomes research and to describe the methods used therein [44]. This review found 98 EHR-based outcomes research studies, with 88% being published in specialty medical journals. Of the outcomes studied, clinical and pharmacologic outcomes were the most common (31% and 19% respectively), whereas economic outcomes were the least common (3%). The study also examined 28 conference abstracts from ISPOR and Academy Health’s Annual Research Meeting. Of these 124 studies, only 78 (63%) used multivariable regression methods to control for confounding, and only 1 study used propensity score methods to control for selection bias. Further, no consideration of handling of missing data in the studies evaluated was given. Finally, this literature review perpetuates the misconception that, “[EHR] data can easily be queried to identify patients based on diagnoses, procedures, and dispensed medications” and that these data are “readily accessible in real time” (p. 618). Such assertions are typical in earlier EHR research literature and are demonstrative of oversimplification of the challenges researchers face when using EHRs for research. As we will see next, more modern papers acknowledge some of these challenges.

In a study that used the biorepositories of five large institutions which are part of the Electronic Medical Records and Genomics Network (eMERGE) [45] for genome-wide

association studies, investigators sought to phenotype diseases [46]. In this study, all sites had different EHRs—including both internally-made and commercially-procured—that were all exceeding meaningful use requirements for EHRs, set forth by the Office of the National Coordinator. However, race, ethnicity, exposure history, and family history of illness had varying rates of capture within the EHR. Further, when captured, these items were often stored in a free-text (structured format), in inconsistent nomenclatures, meaning it had to be parsed out with natural language processing [46]. It is worth re-emphasizing these sites that are part of the eMERGE network, in spite of the challenges presented by EHRs have found 48 disease or condition phenotypes to date [47]. While many of these phenotypes rely on natural language processing, some do not.

In spite of the challenges inherent within the scope of using EHR data for research, the field of EHR data research is still promising. It is just not the easy, golden egg researchers once thought. Other problems still persist, such as censoring, missing data, and attrition. The United States is still a long ways off from having a comprehensive, single medical record for each patient—which some thought simply hinged on increased Internet bandwidth and financial incentives [48]. Even once we have overcome challenges inherent within EHR records, more challenges are systematic due to our system of healthcare delivery, which is highlighted by fragmented care. Unless a patient is seen solely within one integrated health system, her records are in many EHRs—including primary care, emergency care (if not in the same system), and specialty care. Nonetheless, while there are many challenges that must be overcome, including dealing with missing data, it is imperative we press forward to solve some of these challenges as the data within the EHR—while not a golden egg—holds promise.

2.4 Clinical Background – Respiratory Failure

Respiratory failure is a syndrome whereby the lungs fail in their primary function of gas exchange; the lungs fail to adequately expel carbon dioxide or oxygenate the blood. Many conditions can result in respiratory failure, such as chronic obstructive pulmonary disease (COPD), cystic fibrosis, pneumonia, emphysema, chronic bronchitis, pulmonary embolism, and stroke [49, 50]. Further, respiratory failure can be a sequela of surgery or trauma.

Respiratory failure is of large concern for medicine, as it is the leading cause of in-hospital death, and the 3rd leading cause of death in the United States [51, 52]. For patients who experience respiratory failure and require the assistance of a ventilator (VDRF), this typically involves admission to the ICU.

In the following sections the epidemiology of respiratory failure will be reviewed, including where it ranks for cause of death in the United States and its prevalence in the community and among the aged in institutionalized settings. Then a brief examination of the etiology of respiratory failure will be undertaken.

2.4.1 Respiratory Failure: Population Statistics

Chronic lower respiratory disease was the 3rd leading cause of death in the United States in 2014 according to the Centers for Disease Control and Prevention (CDC), claiming 147,101 lives and comprising 5.6% of all deaths [52]. Further, acute respiratory distress was the 8th leading cause of death among newborns, claiming 460 lives and comprising 2.0% of all newborn deaths [52]. In 2010, of the more than 700,000 people who died as an inpatient (2% of all admissions), respiratory failure was the leading first-listed diagnosis (16.5% of deaths), followed by septicemia (16.3%), and pneumonitis due to solids or liquids (13.6%) [51]. Further, the mortality rate among adults has been shown to steadily increase with age, with nonagenarians—those in their 90s—experiencing a nearly four-fold rate of mortality when compared with adults

aged 18-40 (83% vs. 21%) [53]. While the in-hospital death rate for adult patients with respiratory failure has been on the decline—25.3% in 2000, 19.3% in 2005, and 16.5% in 2010—it has been the leading first diagnosis among in-hospital deaths during this period, and the trend is reflective of the overall trend in the decline of in-hospital deaths [51].

A 2010 national survey of assisted living and similar residential care facilities found 4.2% had asthma, 2.0% chronic bronchitis, 10.8% COPD, and 1.2% emphysema—all conditions that could lead to respiratory failure [54]. Of this same population, nearly one-quarter (23.8%) had one or more overnight inpatient stays in a hospital in the 12 months prior.

An examination of the costs of patients who ventilator-dependent using Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS) data using 2009 data showed the costs per ventilated patient varied widely [53]. The highest costs per patient were seen among surviving pre-term infants, with those aged 24 weeks or younger having a median cost around \$200,000. The median costs per patient among adult patients was fairly steady, ranging from \$17,000 to \$25,000 depending on the age group. In nearly all age groups, a similar amount of per-capita money was spent on surviving and non-surviving ventilated patients.

2.4.2 Ventilator-Dependent Respiratory Failure

A prospective study was conducted in 2008 of 60 pediatric ICUs (PICUs) in 13 countries of all children admitted to the PICU in a one-month period during the season when acute lower-respiratory infections were more prevalent in each respective country [55]. This study found that 50.1% of admissions to the PICU required invasive mechanical ventilation, either intubation or tracheotomy. Further, patients who required reintubation following planned extubation was 24%, with the mortality rate being higher amongst patients who required reintubation (21% vs. 1%). For ventilated patients overall in the ICU, the mortality rate estimates vary—ranging from 20-31% in the adult population [56-58] and 13% in the pediatric population [55].

2.4.3 Respiratory Failure: Pathology

Respiratory failure is a syndrome whereby the lungs fail in their primary function of gas exchange; the lungs fail to adequately expel carbon dioxide or oxygenate the blood. Respiratory failure is diagnosed through arterial blood gas measurements [49]. Respiratory failure is classified as *hypercapnic*, meaning the level of CO₂ in the arterial blood is excessive ($P_{aCO_2} > 45$ mm Hg), or *hypoxemic*, meaning the level of oxygen in the arterial blood is inadequate ($P_{aO_2} < 55$ mm Hg when the fraction of inspired oxygen $[FiO_2] \geq 0.60$ mm Hg) [49]. It is not uncommon for respiratory failure to be both hypercapnic and hypoxemic.

2.5 Severity of Illness Scoring

There are a number of severity of illness scoring systems in use in the ICU, including the Acute Physiology and Chronic Health Evaluation (APACHE), Logistic Organ Dysfunction System (LODS), Mortality Prediction Model (MPM), Simplified Acute Physiology Score (SAPS), and Sequential Organ Failure Assessment (SOFA). These ICU scoring systems are widely used for outcome prediction (most commonly mortality), severity of illness stratification—both in clinical trials and research that uses administrative data—and as a case-mix adjustment for comparing quality of care [11].

Severity of illness scores are different, however, from comorbidity scores—such as the Charlson and Elixhauser scores. The Charlson comorbidity score was created to estimate the 1-year mortality risk from comorbidities in longitudinal studies using information manually extracted from inpatient medical record review [59]. The Charlson comorbidity score was modified by Deyo et al. to allow for use with administrative databases by mapping International Classification of Diseases, 9th revision, clinical modification (ICD-9-CM) diagnosis codes to the diseases described by Charlson et al., listing seventeen diagnostic categories, each containing multiple ICD-9-CM diagnoses [60]. The Elixhauser comorbidity score was created specifically

for use with administrative data, giving a set of 30 comorbidities for which a researcher can use for statistical control in multivariable analyses to predict charges, length of stay, or in-hospital mortality [61]. Severity of illness scores differ from comorbidity scores, however, in that they measure physiological derangement—using clinical and laboratory data—rather than presence or absence of comorbid conditions associated with death. As such, they are used in two different manners for risk adjustment, with one adjusting for baseline health and the other adjusting for severity of illness. To be certain, two patients in the ICU with the same comorbidities and Charlson score could have markedly different severity of illness scores, and therefore prognosis.

Further, while both the Charlson and Elixhauser comorbidity scores have been shown to have good predictive ability of mortality among ICU patients—Charlson had 65% area under the Receiver Operating Characteristic Curve (AUROC) and Elixhauser 66% AUROC for 30-day mortality among ICU patients [62]—their predictive ability is likely to be lower than that of severity of illness based scores due to using patient comorbidities, rather than severity of illness scores which measure actual physiological derangement and are more real-time. One study examined the predictive ability of the Charlson score to that of one severity of illness score (SAPS II), finding the severity of illness-based SAPS II score to be superior in prediction of 30-day mortality than the Charlson score at $\alpha=0.05$, 0.821 vs. 0.607 AUROC respectively [63]. Therefore, when available, the researcher is wise to use physiology-based severity of illness scores in addition to comorbidity scores for baseline risk adjustment. For this dissertation, focus will be given to one of the more commonly-used physiology-based severity of illness scoring systems, the SOFA score.

2.5.1 SOFA Score

The Sequential Organ Failure Assessment (SOFA) score was created by the European Society of Intensive Care Medicine (ESICM) in 1994 via a consensus meeting—essentially using

the Delphi technique—to provide an objective scale to describe the degree of organ dysfunction or failure [64]. The score uses information that is routinely-collected in the ICU, making it a scoring system that is easily implemented. The SOFA score was envisioned to have two applications [64]. The first application was to understand the course of organ dysfunction (and failure), including the relationship of multiple organ failure. The second application was as an instrument to be used for baseline severity assessment and measurement of the effects of interventions. The authors of the SOFA score emphatically asserted that it was designed as a tool for description, not prediction [64]. However, its usage has changed over time as it has demonstrated to be a good prognostic tool among ICU patients. The SOFA score is recommended for assessment of septic patients by the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) guidelines [65] and is even used as an element of consideration in emergency triage in some states' crisis standard of care plans [66].

The SOFA score ranges from 0 to 24, composed of 6 sub-scores. The sub-scores have a range of 0 to 4 points being assigned to each of six organ systems: respiratory, hematologic, hepatic, cardiac, neurologic, and renal. A higher score represents a higher level of dysfunction, and thus greater severity. Calculation of the SOFA score is shown in *Table 2.1*, representing the 6 organ systems covered.

Table 1 SOFA score calculation

<i>System</i>	<i>Assigned Score</i>				
	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
Respiration PaO ₂ /FiO ₂ , mmHg	≥ 400	< 400	< 300	< 200 ^a	< 100 ^a
Coagulation Platelets x 10 ³ /mm ³	≥ 150	< 150	< 100	< 50	< 20
Hepatic Bilirubin, mg/dl (μmol/l)	< 1.2 (< 20)	1.2 – 1.9 (20 – 32)	2.0 – 5.9 (33 – 101)	6.0 – 11.9 (102 – 204)	≥ 12.0 (> 204)
Cardiovascular Hypotension	MAP ≥ 70 mm Hg	MAP < 70 mm Hg	Dopamine ≤ 5 or Dobutamine (any dose) ^b	Dopamine > 5 or Epinephrine ≤ 0.1 or Norepinephrine ≤ 0.1 or Phenylephrine ^c ≤ 0.22	Dopamine > 15 or Epinephrine > 0.1 or Norepinephrine > 0.1 or Phenylephrine ^c > 0.22
Central Nervous System Glasgow Coma Scale Score	15	13 – 14	10 – 12	6 – 9	< 6
Renal Creatinine, mg/dl (μmol/l)	< 1.2 (< 110)	1.2 – 1.9 (110 – 170)	2.0 – 3.4 (171 – 299)	3.5 – 4.9 (300 – 440)	> 5.0 (> 440)
Urine output, ml/day				< 500	< 200

^a With respiratory support

^b Administered for at least 1 hour

^c Phenylephrine added by Knox et al. to list of vasopressors according to standard equivalency [67]

The first organ system is respiratory dysfunction, which is calculated as the ratio of the partial pressure of oxygen (PaO₂) to the fraction of inspired oxygen (FiO₂)—often referred to as the *Carrico index* or *P/F ratio* [50]. The partial pressure of oxygen (PaO₂) measures the level of oxygenation within the arterial blood, with normal values ranging from 70-95 mm Hg [68]. The fraction of inspired oxygen (FiO₂) measures the percentage of oxygen in the air being inhaled. This P/F ratio then measures the degree of hypoxemia, with the scores of 2, 3, and 4 matching the

mild, moderate, and severe categories respectively of the Berlin definition of acute respiratory distress syndrome [69].

The second organ system measured by the SOFA score is coagulation, measured as platelet count, where decreasing levels of platelet counts confer a higher coagulation score in the SOFA system. Platelet counts considered to thrombocytopenic—less than 150,000/mm³ [70]—are assigned a score of at least one, with lower counts garnering a higher coagulation component SOFA score.

The third organ system measured by the SOFA score is hepatic, measured as the concentration of bilirubin in the blood. Elevation of serum bilirubin, known as hyperbilirubinemia, occurs when the liver fails to adequately metabolize bilirubin—a byproduct of the metabolism of heme, which is approximately 70-90% hemoglobin of erythrocytes (red blood cells) [71]. Hyperbilirubinemia is typically caused by liver dysfunction or disease, bilirubin metabolism disorders (such as Gilbert syndrome), or biliary tract obstructions [57]. Further, elevated serum bilirubin levels have been shown to be predictive of short-term mortality [72-75]. Increasing amounts of bilirubin correspond to a higher hepatic component SOFA score.

The fourth organ system measured by the SOFA score is cardiovascular, measuring hypotension and pharmaceuticals administered to return the patient to a normotensive state. A patient is assigned a score of 1 when deemed hypotensive, defined as a mean arterial pressure (MAP) < 70 mm Hg. Mean arterial pressure is calculated as follows,

$$MAP = \text{diastolic blood pressure} + \frac{\text{systolic} - \text{diastolic}}{3}$$

and a hypotensive state means the body's organs are being insufficiently perfused [76]. As the patient exhibits greater hypotension, increased amounts of vasopressive drugs are administered—such as dopamine, dobutamine, epinephrine, or norepinephrine—to constrict the blood vessels,

with the goal of raising the MAP and returning the patient to a normotensive state [76].

Increasing doses of vasopressors correspond to a higher cardiovascular component SOFA score.

The fifth organ system measured by the SOFA score is the central nervous system, as measured by the Glasgow Coma Scale. The Glasgow Coma Scale (GCS) was developed in 1974 by two physicians to quantify the level of consciousness of critically ill patients by measuring three aspects of behavior: eye, verbal, and motor response to allow for longitudinal monitoring [77]. The GCS is scored as shown below, from 3 to 15—with a lower composite score representing a worse prognosis. Each category is rated as the best response for the category that uses a standardized approach for evaluation.

Table 2 Scoring of the Glasgow Coma Scale (GCS)

<i>Eye Response</i>	<i>Verbal Response</i>	<i>Motor Response</i>
1. None	1. None	1. None
2. Open to pain	2. Incomprehensible speech	2. Extension
3. Open to speech	3. Inappropriate speech	3. Abnormal flexion
4. Open spontaneously	4. Confused conversation	4. Normal flexion (withdrawal)
	5. Orientated	5. Localizing response
		6. Obeys commands

The GCS has been integrated into intensive care medicine in over 80 countries, with the three components being used to describe the impairment of consciousness on individual patients [78]. Further, the GCS is used as a risk-adjustment or prognostic tool in outcomes research [78]. Finally, a demonstration of the clinical importance of the GCS is its incorporation into the International Classification of Diseases, 10th revision, Clinical Modification (ICD-10-CM) which allows for component and composite GCS scores to be coded (code R40.2xx), along with the time of measurement [79].

The sixth, and final, organ system measured by the SOFA score is renal, measured as creatinine clearance or daily urine output. An elevated serum creatinine or decreased urine output

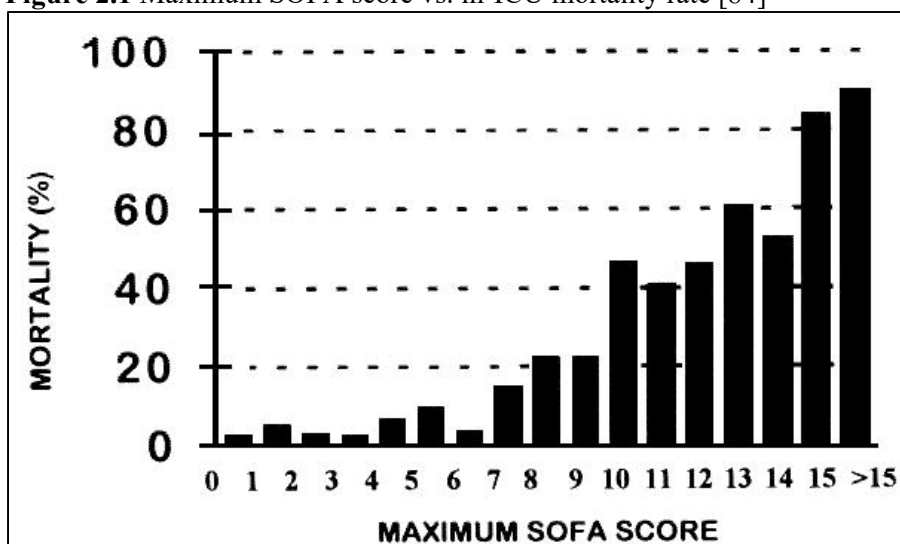
are signs of diminished kidney function, possibly acute kidney injury [80]. Worldwide the incidence of acute kidney injury was estimated at 21.6% in hospitalized adults and 33.7% in hospitalized children, with mortality rates of 23.9% and 13.8% respectively in a meta-analysis of 154 studies of 3.6 million people [81]. Further, two classification systems of acute kidney injury—which involve change in serum creatinine levels and daily urine output—have been shown to be predictive of outcomes in ICU patients, including mortality, renal failure, and length of stay [82, 83].

2.5.2 Interpretation of SOFA Score

According to the Sepsis-3 consensus paper, the baseline SOFA score—which is calculated upon admission to a critical care unit—should be assumed to be zero, unless the patient has a known organ dysfunction [65]. A change in total SOFA score of at least two points represents organ dysfunction, and a SOFA score that is 2 or greater is associated with a 10% in-hospital mortality rate [65].

The SOFA score was validated in an ICU setting, which demonstrated the presence of sepsis was associated with higher component organ SOFA scores [84]. Further, of the patients whose ICU stay was for at least 7 days, an increase of SOFA scores from baseline was correlated with greater odds of death (44% of non-survivors vs. 20% of survivors; $p < .001$); whereas a decrease in SOFA score from baseline was correlated with greater odds of survival (21% of non-survivors vs. 33% of survivors; $p < .001$) [84]. When examined as the maximum SOFA score throughout an ICU admission, there is a trend toward increasing in-ICU mortality as the maximum SOFA score increased, as shown below in Figure 2.1 (below).

Figure 2.1 Maximum SOFA score vs. in-ICU mortality rate [84]



The SOFA score has been examined for other clinical populations as well. The SOFA score has been modified and adapted for the pediatric population (pSOFA) in predicting in-hospital mortality, showing excellent discrimination of 0.94 AUROC (95% CI: 0.92-0.95) using age-adjusted SOFA parameters [85]. Additionally, the SOFA score on the day of admission to the ICU has been compared to the APACHE II score in oncology patients admitted to the ICU and found superior at predicting in-ICU mortality (0.925 AUROC, 95% CI: 0.859-0.991 vs. 0.710, 95% CI: 0.578-0.843 respectively), and similar in predicting in-hospital mortality (0.835 AUROC, 95% CI: 0.734-0.934 vs. 0.655, 95% CI: 0.491-0.819) [86]. The SOFA score on the seventh day post-transplant has also been shown to be highly predictive of mortality for living-donor liver transplant recipients in predicting 3-month post-operative mortality (0.952 AUROC, 95% CI: 0.874-1.00) [87]. Similarly, the SOFA score has been used to predict mortality among trauma patients [88], those with hematological malignancies [89], patients in acute geriatric care settings [90], and even ICU-treated refractory status epilepticus patients [91]. Finally, the

admission SOFA score has also been shown to be associated with diminished quality of life one year post-discharge among ICU survivors, as measured by the EuroQoL-5D [92].

2.5.3 Examples of SOFA Score in the Literature

The SOFA score is often used in outcomes studies of ICU-treated conditions as a predictor or adjustment variable. The SOFA score serves as a measure of patient severity within the ICU, which is a measure upon which one can statistically control to examine outcomes.

One prospective, multicenter study of adult ventilated ICU patients examined the risk of developing adult respiratory distress syndrome (ARDS) [93]. In the final statistical model that examined risk of ARDS, the baseline SOFA score—along with other covariates, such as BMI and functional status—were used to predict risk of developing ARDS among ICU patients. However, the study admitted that data were missing for elements of the SOFA score, with those cases simply being omitted from the analysis. This is problematic, as the number of cases excluded was not mentioned and the potential for biased findings due to the missingness was not discussed.

2.5.4 Missingness of SOFA Score Items

Missingness of SOFA score items, or the total score itself, varies across studies. In one study the admission SOFA score had 0% missingness, but was a prospective one-year study [92].

In the validation study for the SOFA score, bilirubin values were the most commonly missing item, whereas platelet counts were the most infrequently missing—however the percent of time these items were missing was not mentioned [84]. For imputation of missing items, the mean of the value prior to and after the missing value was imputed; in cases where multiple observations were missing, the missing value was left untouched—resulting in available case analysis being used for the analyses [84]. In a later study by some of the same authors as the validation study, similar missingness was found, with bilirubin being the most commonly missing item, and platelets and PaO₂/FiO₂ ratios being the most infrequently missing items [94].

2.6 Mechanisms of Missingness

The mechanism of missingness is the process that governs whether data are missing. The theory for missing data mechanisms was first proposed by Rubin—applying it to survey designs [95], which was simplified by Little & Rubin [96] to remove the response parameter that is inherent in survey methodology. Using the groundwork laid therein, the mechanisms of missingness can be defined symbolically.

If a complete dataset is specified as $Y = (y_{ij})$ with i representing subjects (as rows) and j representing variables (as columns), with a size of $(n \times K)$. Then y_i is the vector of variables for subject i , which can be expressed as $y_i = (y_{i1}, \dots, y_{iK})$. This complete dataset is shown below in *Figure 2.1*, which shows a matrix Y with a size of (3×4) , representing 3 observations—each observation containing 4 variables. In this table you can see cell $y_{3,4}$, which represents the 4th variable for the 3rd subject.

Figure 2.2 Complete data matrix

		<i>j</i> Variables			
		1	2	3	4
<i>i</i> Subjects	1				
	2				
	3				$y_{3,4}$

To indicate whether data are missing, $M = (m_{ij})$ is a matrix of binary variables of the same size as Y , with $m_{ij} = 1$ indicating the datum at y_{ij} is missing and $m_{ij} = 0$ indicating the datum at y_{ij} is observed. The mechanism of missingness is represented by a conditional

distribution of M , where $f(M|Y, \phi)$, where ϕ are latent parameters. And thus, the groundwork of introducing the symbols for defining the missing data mechanisms has been laid.

2.6.1 Missing Completely at Random (MCAR)

The first missing data mechanism is known as *Missing Completely At Random* (MCAR), meaning the probability of missingness does not depend values of Y , whether observed (Y_{obs}) or missing (Y_{miss}). Using the Little & Rubin equation, MCAR can be written as

$$f(M|Y, \phi) = f(M|\phi)$$

for all values of Y and ϕ . In the case of MCAR, when data are missing it is equivalent to a simple random sample of the full dataset [97]. This is equivalent to asserting that the function of missingness cannot be described by the data, but rather is a stochastic process modeled by the latent parameters ϕ . Further, the probability of missingness for one variable can be related to the probability of missingness of another variable, such as is the case of unit non-response in a survey [97]. The MCAR equation given above has been made more comprehensible by Allison [97]—who adds X as a vector of fully-observed variables, writing MCAR as

$$\Pr(Y_{miss}|Y, X) = \Pr(Y_{miss})$$

The MCAR mechanism is akin to taking a completely random sample (Y_{obs}) from a population (Y). Further, excluding Y_{miss} from any analysis should not bias the results of an analysis at moderate percentages of missingness.

2.6.2 Missing at Random (MAR)

The second missing data mechanism is known as *Missing At Random* (MAR), which means that the probability of missingness does not depend on the missing values of Y , (Y_{miss}), but may depend on the observed values of Y , (Y_{obs}). The MAR equation can be written as

$$f(M|Y, \phi) = f(M|Y_{obs}, \phi)$$

for all values of Y_{miss} and ϕ . Once again Allison makes this equation more comprehensible, representing it as

$$\Pr(Y_{miss}|Y, X) = \Pr(Y_{miss}|X)$$

MAR means that the missing values do not depend on the values themselves, but can depend on other observed values [98]. Some authors assert that it is possible to take data from NMAR to MAR in survey and intervention designs through asking subjects how likely they are to drop out of a study, or by using proxy measures that are highly correlated with the missing covariate [99], thus adding a predictor of missingness to the Y_{obs} vector.

2.6.3 Missing Not at Random (MNAR)

The third, and final, missing data mechanism is known as *Missing Not At Random* (MNAR), which means that the probability of missingness depends on the missing values of Y , (Y_{miss}) themselves. The MNAR equation can be written as

$$f(M|Y, \phi) = f(M|Y_{miss}, \phi)$$

for all values of Y and ϕ . An example of MNAR is found in job applications where a question asking if a person has ever been arrested would depend on the answer itself. A person who has never been arrested will readily answer *no*, however one who has been arrested would be apprehensive to answer that question. Another example of MNAR from the literature is the issue of non-response of income reporting to the U.S. Census Bureau, where those with higher incomes were less likely to report their incomes [100].

The importance of understanding the missing data mechanism—whether MCAR, MAR, or MNAR—has been covered well in the literature [19, 95, 98]. Several studies have suggested less discrete definitions of these mechanisms. One study posits a more fluid definition of missingness, asserting it is more like a continuum between MCAR and MNAR [101], although interesting, the point is quixotically impractical. Another theory-building simulation study that examined missing

data methods using all three missing data mechanisms asserted one single mechanism is unlikely to be the cause of missingness [102]. The focus must remain on having a solid understanding of the missing data process, as selecting an analytical approach to missing data is predicated on an understanding of the missing data mechanism.

2.6.4 Tests for Missing Data Mechanism

Unfortunately, few tests exist to distinguish between the three missing data mechanisms and the ones that do exist have their limitations. The most commonly-cited test is known as *Little's test*, which gives a single χ^2 test statistic for testing the MCAR assumption on multivariate continuous data [103]. It is useful for situations where the researcher is trying to ascertain if the data are MCAR or MAR. Essentially, Little observed that one would need to split the data into two groups, those with missing values for a given variable and those without missing values. Then one would compare the distributions for each variable in the dataset between these two groups using a two-sample Student's *t*-test, with a significant difference indicating the data are not MCAR. However, for p number of variables in a dataset this would yield $p(p - 1)$ *t*-tests. To get around the multiple comparisons, Little created his likelihood ratio test that is asymptotically chi-squared. However, situations where missingness is due to the MCAR mechanism are rare [97, 104], and categorical data are common, limiting the utility of this test.

Other tests aimed at longitudinal data include ones developed by Park & Davis, which is an extension of Little's test but for repeated measures categorical data that tests for MCAR [105], and a test by Park & Lee based on generalized estimating equations that uses a missing indicator using a pattern-mixture model [106].

Heitjan & Basu examined the MCAR and MAR missing data mechanisms from both the Bayesian and Frequentist statistical inference perspective showing different ignorability requirements exist through statistical simulation [107]. For Bayesian inference—such as would be

done through using maximum likelihood estimation (MLE) to deal with missing data—showing the ignorability condition is met if the likelihood function is the same with or without accounting for the missing data mechanism. However, for Frequentist inference—such as would be done through multiple imputation—the ignorability condition is only met if the data are MCAR.

2.7 Missing Data Patterns

The pattern of missing data is important to understand, as various methods require a certain data pattern. The pattern of missing data is essentially the unique patterns of values of the missing data vectors, m_i , in the missing data matrix $M = (m_{ij})$. Recall that i represents subjects (as rows) and j represents variables (as columns). These patterns of missing data vectors are arranged from most number of observed variables to least number of observed variables, as shown in Figure 2.3 below.

Figure 2.3 Missing data patterns, monotonic vs. non-monotonic

<i>Monotonic</i>				<i>Non-Monotonic</i>			
V ₁	V ₂	V ₃	n	V ₁	V ₂	V ₃	n
X	X	X	100	X	X	X	100
X	X	•	50	X	•	X	50
X	•	•	30	X	X	•	30

In Figure 2.3 one can see columns of three variables and the number of observations with that pattern. In the cells below an X represents an observed value (Y_{obs}), whereas a dot represents a missing value (Y_{miss}). A monotonic missing data pattern is one that will follow a stepwise fashion as shown, where once a variable is missing in the list of patterns, that variable will always be missing in subsequent patterns (c.f. Figure 2.3, V₂). By definition, a dataset which has only

one missing variable will be necessarily monotonic. Further, when only one variable is missing, analytical methods to address the missing data are known as *univariate* methods—regardless of whether the imputed variable is a dependent or independent variable in later analyses; when more than one variable is missing, analytical methods to address the missing data are known as *multivariate* methods.

2.8 Amount of Missing Data

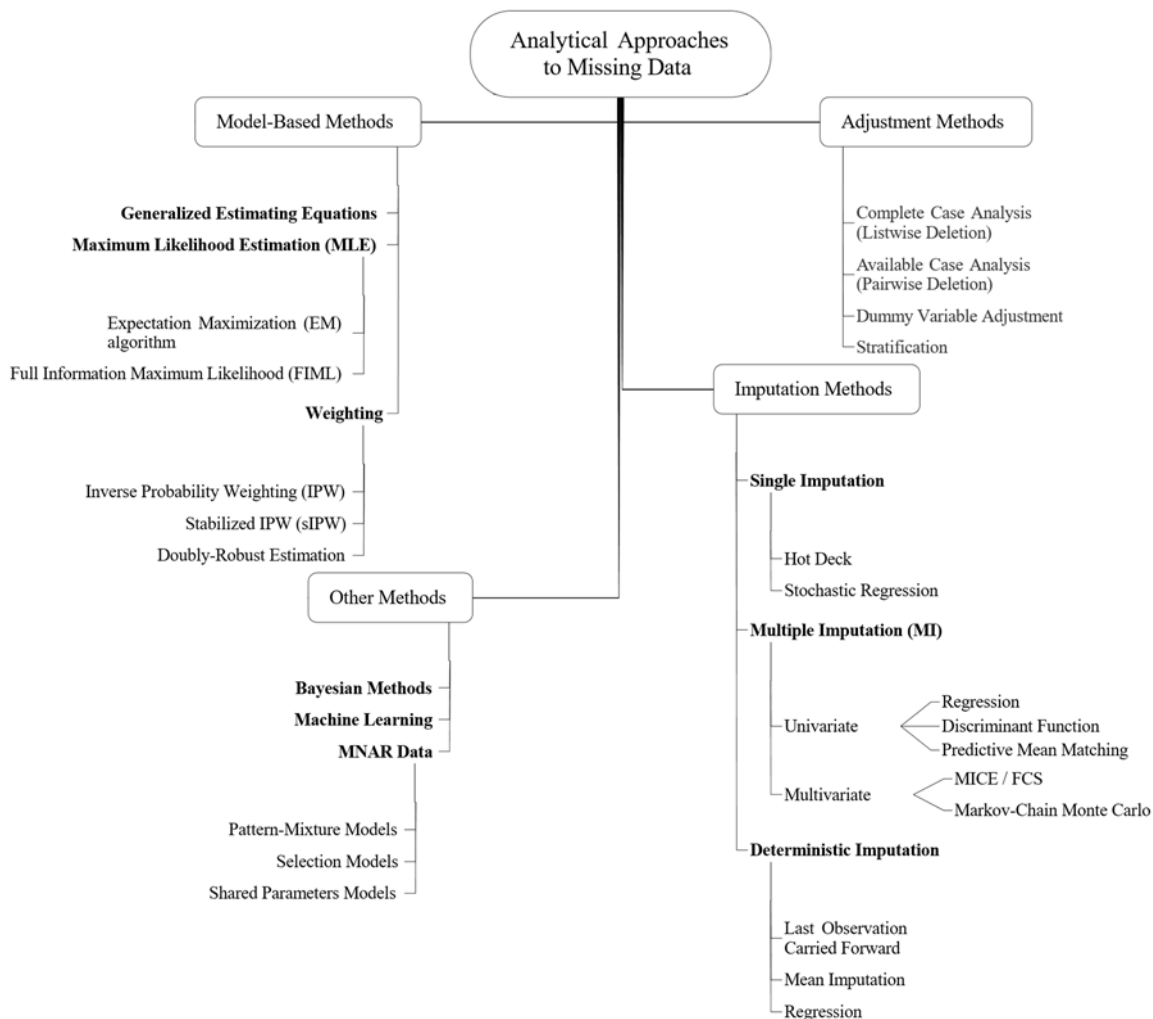
The amount of missing data, represented symbolically by γ , is of concern in all statistical analyses due to risk of bias. The amount of missing data that is permissible prior to unacceptably biasing results is not well-established. The general guideline in the missing data literature is $\gamma \leq 5\%$ is considered trivial [12]. However, for clinical trials data Yeatts and Martin caution the range of 5-20% is of most concern because the rate is high enough to cause statistical bias, yet not high enough for the findings to be rejected solely on the basis of missingness [108].

2.9 Analytical Approaches to Missing Data

How to handle missing data is a subject that has been explored deeply in statistical literature; a taxonomy of analytical approaches to missing data is given below in Figure 2.4. Moreover, strategies may vary depending on whether the missing data are outcomes or covariates; the discussion henceforth centers on missing covariate data.

In addition to CCA—which relies on MCAR—there are other more sophisticated techniques, all of which depend on the missing data mechanism. Alternatively, there exists multiple imputation (MI) techniques which allow for multiple datasets to be created with various values imputed for each missing value when the MAR mechanism for missing data is likely. Each dataset is analyzed separately, then the parameter estimates and confidence intervals from each separate analysis combined using Rubin's rules [98]. MI techniques allow for the uncertainty surrounding the values of the missing data to be accounted for in the analysis.

Figure 2.4 Taxonomy of analytical approaches to missing data



2.9.1 Adjustment Methods

Missing data are a problem in research, as statistical software rely on complete data for all variables in a regression model. If one value for a variable is missing the entire observation is excluded from the analysis, which leads to problems of potential bias and reduced power due to smaller sample size. There are four main adjustment methods for dealing with missing values: complete case analysis, available case analysis, dummy variable adjustment, and missing data stratification.

The first adjustment method is known as complete case analysis (CCA), or listwise deletion. With this method the researcher removes all observations that have a missing value for any of the outcome or explanatory variables, for all intended analyses.

The second adjustment method is known as available case analysis (ACA), or pairwise deletion—the default method of all statistical software. With this method any observation that has a missing outcome or explanatory variable is excluded from the analysis. This leads to problems of changing samples, as during model fitting as potential covariates are added and removed the sample size will change. Further, secondary analyses will also have different sample sizes—and essentially different samples—than the primary analysis. Clearly, such a strategy is problematic as will be discussed next.

The first problem with CCA and ACA methods is that reduced statistical power is achieved due to a smaller sample size. While the magnitude of this problem varies based on the total sample size and the percent of missing data, it is still a problem. The second problem with CCA and ACA is that unless the data that are missing are missing completely at random (MCAR), the results will likely be biased [109]. If it is completely a chance occurrence the data are missing—or MCAR—removing the case will not bias the results using CCA [109]. However, CCA and ACA will result in a loss of precision of the estimate, yielding a wider confidence interval. Therefore, dealing with the missing data, rather than ignoring it, is warranted.

The third adjustment method available is dummy variable adjustment. With this method, a constant value is imputed for all missing values—often the mean of the observed data—when a value is missing for a predictor in a regression model [110]. Then a missing data indicator is added, such that 1 indicates the datum was missing and 0 indicates the datum was observed. In the analysis both the predictor with the missing and imputed data along with the

indicator are regressed on the outcome of interest. This model has the prima facie advantage of using all available data. However, dummy variable adjustment has been shown to create severely biased parameter estimates—both in magnitude and direction and often in an unpredictable direction—even when data are MCAR or the percent of missing data is low (e.g. 2.5%) [109, 111].

The fourth method available to the researcher is missing data stratification. This method is common in health services research when categorical data are missing, such as race or marital status. In this method an additional *missing* stratum is created. As was the case with the dummy variable adjustment method, missing data stratification has the prima facie advantage of using all available data. However, despite the fact that this results in severely biased parameter estimates [111], it is still of common use [112].

2.9.2 Imputation Methods

There are two classes of imputation techniques. The first class is deterministic imputation, which fills-in—or imputes—one value for every missing value. The second class of imputation techniques incorporates randomness into the imputed values, with two variants of this class. The first variant imputes one value for every missing value, which is known as single imputation. The second variant is known as multiple imputation (MI), which also imputes values, but does so a number of times—creating multiple datasets. These datasets are then analyzed using normal analysis techniques, whereby the point estimates and standard errors are combined using standard combining rules known as *Rubin's rules* [98].

2.9.2.1 Deterministic Imputation

There are at least three variants of deterministic imputation methods: last observation carried forward (LOCF), mean imputation, and regression imputation. All of the deterministic

imputation techniques create one complete dataset by filling in the missing values to allow all the data to be used.

The first deterministic imputation method is known as last observation carried forward, which is used in repeated measures data. The researcher simply imputes the last observation's value for all subsequent missing observations, until (and if) a subsequent observation is recorded. Unfortunately, in longitudinal data one must be concerned about the mechanism of missingness, as there is likely a systematic difference between those who complete a study and those who do not, likely making these data MAR or MNAR.

The second deterministic imputation method is known as mean imputation. Here the researcher simply imputes the mean value, or median in the case of skewed data, value for all missing continuous data. This method has been rejected by Rubin as being unacceptable for research [113], and has been repeatedly shown in studies to introduce unacceptable bias and over-precise confidence intervals [114] because it artificially reduces overall variance.

The third, and final, deterministic imputation method is known as regression imputation. Here the researcher uses linear regression to predict missing values using the complete cases. For instance, if there are three variables in our dataset (X_1, X_2, X_3) with X_3 containing missing values, one would regress X_3 on X_1, X_2 . This would yield $E(X_3) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$, allowing the researcher to impute the missing values of X_3 .

Deterministic imputation techniques allow for standard statistical techniques to be used, as if no data were missing. However, these methods suffer from various drawbacks—including biased estimates, and naively-small confidence intervals of those estimates, as the uncertainty surrounding the missing data is not accounted for in the analysis [98, 113, 115].

2.9.2.2 Single Imputation

Similar to the deterministic imputation methods, single imputation methods create a complete dataset by filling in the missing values with plausible values to allow all the data to be used—rather than discarded, as is the case with CCA. However, single imputation methods differ from the deterministic methods in that they introduce random variation into the imputed values. There are at least two main variants of single imputation: hot deck imputation, and regression imputation with a random component.

The first single imputation method, known as hot deck imputation, was developed at the U.S. Census Bureau in the 1960s to address survey item non-response to the question of household income in the Current Population Survey [100]. Essentially, the hot deck method finds all the observations in the dataset that are similar to the observation with a missing variable—the *hot deck*—then randomly picks one of the observations from the group of similar fully-observed observations and imputes that value into the observation with missing data.

To illustrate hot deck imputation, suppose income is missing for a Caucasian male, 33 years of age, who worked full-time in a professional occupation, in the state of Florida. Hot deck imputation would find all males in the age group 25-34, who are full-time professionals in the Southeastern United States. Then, a random pick of one person from this group of fully-observed data would be made, and the income of this individual being imputed into the missing observation.

Advantages of hot deck imputation include that it makes a logical assumption regarding imputed values, as one would expect a missing value to be very similar to a nearly identical person's value. Another advantage is that hot deck imputation imputes realistic values for missing observations, since the imputed values are drawn from the fully-observed dataset.

A number of adaptations of hot deck imputation exist, including predictive mean matching (PMM) which uses linear regression to estimate the values of missing continuous data, then randomly picks one of the similar case's values to use as the imputed value [116, 117]. Interestingly, the hot deck imputation method is still being used by the U.S. Census Bureau today [118, 119].

Another single imputation method is known as regression imputation with a random component, also referred to as stochastic regression imputation. Here the method is nearly identical to the deterministic regression imputation technique, however a random component is introduced. Using the previous example with three variables in our dataset (X_1, X_2, X_3) where X_3 contains missing values, one would again regress X_3 on X_1, X_2 . This would yield $E(X_3) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$. However, one would then introduce randomness by adding a random component through multiplying the error term ε (which is the root mean squared error of the regression equation) times a random draw from the standard normal distribution ($\sim N(0,1)$). This regression equation with the random component added then allows the researcher to impute the missing values of X_3 .

Single imputation methods suffer from the same drawbacks as deterministic imputation techniques, such as biased estimates and small confidence intervals of those estimates, and the uncertainty surrounding the missing data is not accounted for in the analysis [98].

2.9.2.3 Multiple Imputation

Multiple imputation (MI) techniques are similar to single imputation techniques, but they create multiple datasets which Rubin (1987, p.2) described as “representing a distribution of possibilities” [98]. Multiple imputation is used when the MAR missing data mechanism is likely. Multiple imputation techniques are also believed to be an option when the data are MNAR, so

long as the missing data mechanism is correctly specified [97]. However, the probability of correctly modeling the missing data mechanism is unlikely to occur.

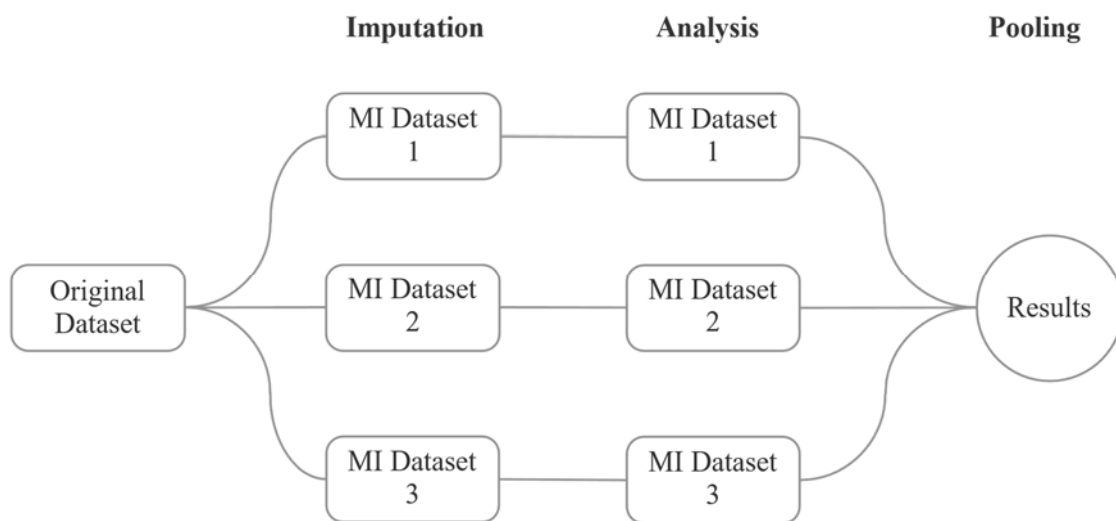
The goal of MI is not to make up data, but rather to allow all the data that are present to be used in analyses to achieve valid statistical inference, not perfect point prediction [113]. In fact, even when the percentage of missing data is small and the data are MCAR—meaning the point estimates generated in an analysis will be unbiased—MI has the advantage of increased power, yielding tighter confidence intervals around those estimates. This was demonstrated in Bounthavong et al. (2015) which compared CCA to MI to study the outcome of dyslipidemia in the Veteran population using electronic medical record (EMR) data [120]. This study found that with 22% missingness the point estimates changed very little, but the confidence intervals were much narrower using MI ($m = 5$), demonstrating the utility of MI even when data are believed to be MCAR.

There are techniques for univariate and multivariate missingness. The techniques for univariate missingness are the same as the single imputation techniques described previously, varying only by creating m number of multiply-imputed datasets. However, prior to explaining specifics of various MI techniques, it is prudent to explain the general process of multiple imputation process.

Essentially, MI fills in the values that are missing with plausible values, to allow all of the existing data to be used—rather than discarded, as is the case with CCA. Multiple imputation is composed of three phases [95, 98], see Figure 2.5 below. The first phase fills in the missing values for each variable through one of several MI techniques, repeating this process a certain number of times (m times). This first phase is where all MI techniques vary. The second phase is where each of the m multiply-imputed datasets are analyzed separately using normal regression analysis techniques, yielding m number of vectors of parameter estimates and standard errors.

The third, and final, phase is where the m vectors of parameter estimates and standard errors are pooled together using a technique commonly known as *Rubin's rules* [98], which essentially averages the repeated parameter estimates and standard errors from the m datasets to give one vector of parameter estimates and another of standard errors.

Figure 2.5 Conceptual diagram of multiple imputation



There are two common methods of MI for multivariate missing data, multivariate imputation through chained equations (MICE) and Markov-Chain Monte Carlo (MCMC) multiple imputation. These two MI methods vary based on their statistical assumptions and execution, which are discussed below.

The first MI method is Multivariate Imputation through Chained Equations (MICE), which also is known as Fully-Conditional Specification (FCS) in the literature [121]. MICE is a special type of MI whereby various regression-based techniques—typically generalized linear models, such as linear or logistic regression—can be used for each type of missing variable in the

data, using the observed variance and covariance matrices. In other words, a logistic regression model can be used to help predict missing binary values; an ordinary least-squares (linear) regression model can be used to help predict missing continuous variables. The MICE method also has further flexibility through allowing discriminant function methods to predict missing binary or nominal data [12], and predictive mean matching (PMM) for continuous data [122]. One model is created for each variable with missing data. Each of these models are used to impute the missing values, with MICE iterating through m times to create m multiply-imputed datasets, which are then combined in the same manner as all other MI models, using *Rubin's rules*.

The amount of missing data, γ , is of concern with any missing data model. The MICE method has demonstrated good results across a range of missingness, from $\gamma = 5$ to 90%. One study by Janssen et al. (2010) compared MICE to other simpler methods—namely CCA and dropping of predictors with missing data [123]. The study examined a wide-range of missingness (from $\gamma = 10$ to 90%) under the MAR missing data mechanism for three predictors of deep venous thrombosis (DVT), using 500 simulated datasets. For the MICE method, $m=10$ multiply-imputed datasets were created. The MICE method worked well at all percentages of missingness, yielding less bias in the regression coefficients for these predictors and better coverage of the 95% confidence interval of the full dataset regression coefficient.

Another theory-building simulation study by Knol et al. (2010) compared MICE to CCA and the missing indicator method (MIM) for dealing with missing observations for one predictor of major depressive disorder (income) using data from a prior clinical trial [109]. The study examined five levels of missingness ($\gamma = 2.5\%$, 5%, 10%, 20%, and 30%) under the MCAR and MAR missing data mechanisms, using 1,000 simulated datasets. For the MICE method, $m=5$ multiply-imputed datasets were created. This study found that both CCA and MIM resulted in

biased results, with the amount of bias increasing with the amount of missing data. The authors recommended that MIM and CCA never be used, even with small percentages of missing data.

Unfortunately, MICE does not always converge to the correct posterior distribution. There is at least one example from the literature where MICE failed to converge in a study that used EHR data, albeit there was 70% missingness [124]. Nonetheless, MICE has advantages in CER studies due to the wide range in the types of variables for which one might control in a regression analysis, allowing separate imputation models to be used for each variable with missing data. Typically, continuous, ordinal, categorical, and binary variables will all be used in the same analysis as covariates—demonstrating the appeal of using MICE.

The second MI method is Markov Chain Monte Carlo MI (MCMC), which is a method of multivariate normal imputation (MVNI) that assumes multivariate normality of all continuous variables [12]. In contrast to the MICE method, which specifies a conditional distribution to predict missing values for each missing data type, the MCMC method specifies a single joint distribution for all variables with continuous data. For continuous variables that are skewed, transformation prior to using the MCMC method can yield good results [125].

2.9.3 Likelihood Methods

Likelihood methods rely on maximum likelihood estimation (MLE) to estimate parameters for predictors that predict an outcome. Maximum likelihood estimation is similar to MI in that a guess is made at the missing values, however it is done in a more implicit—rather than explicit—manner [126]. The manner in which this is accomplished is through finding the parameter estimates that would maximize the probability of observing what has been observed, known as the maximum likelihood method [115]. Maximum likelihood is also the method by which generalized linear models are solved, whereby the difference between the observed and predicted data is minimized. For one familiar with calculus, this is akin to finding the maxima of

a function by conducting a second derivative test; in this case, the likelihood function is maximized to find the point of highest probability. Upon completion of MLE, parameter estimates will have been made, along with standard error estimates, without the need of creating multiple datasets.

Maximum likelihood estimation is used when the MAR missing data mechanism is likely. However, MLE methods are also believed to be an option when the data are MNAR, so long as the missing data mechanism is correctly specified [115]. Unfortunately, MLE methods rely on the assumption of multivariate normality—meaning that all continuous variables are normally distributed and can be defined as a linear function of all the other variables, with the error terms having equal variance (homoscedastic) and a mean of zero—a very strong assumption [104].

2.9.3.1 Maximum Likelihood Estimation with EM Algorithm

The first type of maximum likelihood method is maximum likelihood estimation using the Estimation-Maximization (EM) algorithm [101, 127]. This method relies on the MAR or MCAR assumption [127] and produces unbiased parameter estimates, but has a drawback in that it does not provide estimates of the standard error for each parameter estimate [104, 115].

There are two steps involved with this method: expectation and maximization [127]. The first step, *Expectation*, imputes values for each missing value in the dataset. Next a regression equation is constructed using the other variables to predict the missing value. Values are then imputed into the dataset for all variables with missing data (Y_{miss}), creating the initial full dataset.

The second step, *Maximization*, involves recalculating the means, variances, and covariances using the dataset from the prior step. When calculating the variances and covariances the residual, or error, term used in the regression equation would be incorporated into these calculations. For example, if Var_1 and Var_2 were used to predict Var_3 , meaning $E(Var_3) =$

$Var_1\beta_1 + Var_1\beta_1 + \epsilon$, then ϵ would help inform the variance and covariance matrix for all Var_3 [104].

2.9.3.2 Full Information Maximum Likelihood

The second type of maximum likelihood method is Full Information Maximum Likelihood estimation (FIML), also known as Direct Maximum Likelihood. One downside to FIML is that it converges more slowly than with MLE using the EM algorithm [127]. However, it is often preferred over MLE using the EM algorithm because it gives accurate estimates of the standard error estimates [104].

2.9.4 Methods Available in Common Statistical Software

Surveying the statistical software for analytical approaches available is predicated on knowing which statistical software packages are used in health services research—unless one’s goal is an exhaustive review of all statistical software. One comprehensive review of statistical software conducted in 1997 identified 220 statistical programs—of which 39 were general statistical (e.g. SAS, SPSS), 25 were for mathematical statistics (e.g. Matlab), 14 for econometrics, and 142 were specialized statistical software for multivariate analyses, specialized modeling, or power and sample size calculation [128]. Keeping to an approach which favors brevity, the literature was consulted to answer this question. Those software with broad representation in health services research journals seems to be SAS (SAS Institute, Inc.; Cary, North Carolina), SPSS (IBM Corp.; Armonk, NY), Stata (StataCorp LLC; College Station, TX), and infrequently RStudio (RStudio Inc.; Boston, MA).

One study in 2014 examined the statistical methods and software used by all studies that used the Canadian community health survey data, published from 2002-2012 (n=663) [129]. This study found the most common statistical software to be SAS (30.8%), followed by Stata (13.1%), SPSS (12.8%), SUDAAN (6.5%), and all others (2.6%). Another study published in 2011

examined statistical software used in health services research published in the United States from 2007-2009 (n=877) [130]. This study found the most common statistical software to be Stata (46.0%), followed by SAS (42.6%), SUDAAN (6.2%), and SPSS (5.8%). However, only 61% of these articles mentioned the statistical software used. Of note, many of the articles mentioned more than one statistical application being used—explaining why the percentages exceed 100%. As one can see, the three most common statistical packages appear to be SAS, SPSS, Stata, and SUDAAN. However, SUDAAN is typically used for more complex research designs—such as correlated, clustered, or stratified designs [131]. The next step is to clearly define the approaches available in these software packages to handling missing data—as they will be the ones most handily available to health services researchers.

A review of the documentation for the latest versions of SAS (v. 9.4; SAS/STAT 14.3; released December 2017), SPSS (v. 25; released August 2017), and Stata (v. 15; released June 2017) was conducted. The results of this review are included below in Table 3.

Table 3 Missing data methods available in SAS, SPSS, and Stata

	SAS	SPSS	Stata
Adjustment Methods			
Listwise Deletion (Complete Case Analysis)	Yes	Yes	Yes
Pairwise Deletion (Available Case Analysis)	Yes	Yes	Yes
Imputation Methods			
Default # datasets	25	5	10
Monotonic Imputation			
Regression - Linear, Logistic, Ordered Logistic	Yes	Yes	Yes
Predictive Mean Matching	Yes	Yes	Yes
Discriminant Function (Nominal data)	Yes	No	No
Propensity Score (Continuous data)	Yes	No	No
Pattern-Mixture Models	Yes	No	No
Non-Monotonic Imputation			
Markov-Chain Monte Carlo - Full dataset	Yes	No	Yes
Markov-Chain Monte Carlo - Monotone	Yes	No	No
Multivariate Imputation through Chained Equations(MICE)	Yes	Yes	Yes
Pattern-Mixture Models	Yes	No	No
Likelihood Methods			
Expectation Maximization (EM) algorithm	Yes	Yes	Yes
Full Information Maximum Likelihood (FIML)	Yes	No	Yes

The default missing data method for a statistical analysis in all three of these programs is listwise deletion (or complete case analysis). All three programs also offer maximum likelihood estimation using the common expectation-maximization algorithm. All the programs also offer multiple imputation (MI), with varying levels of features—which are discussed next.

SAS offers the most methods for addressing both monotonic and non-monotonic missingness patterns, offering all the major MI methods. While Stata allows for MCMC MI, it only does so as a full imputation of the dataset for every missing variable. SAS, however, allows the researcher to use MCMC MI to move the dataset from a non-monotonic to monotonic missing

data pattern—imputing data for variables until a monotonic missing data pattern is reached. Neither SPSS nor Stata allow the researcher to use Discriminant Function nor Propensity Score methods for MI for monotonic missingness. However, it is worth mentioning these two methods are possible through some statistical programming as SPSS and Stata have built-in methods for discriminant function analysis and propensity score methods. Finally, the default number of datasets constructed with multiple imputation, m , varies greatly by software program—from 5 in SPSS, to 10 in Stata, and 25 in SAS (which was changed from 5 with SAS/STAT 14.2).

This section has shown for non-monotonic missingness there are two major options for the researcher who decides not to ignore the missing data: MICE or MLE. With monotonic missingness, such as when only one predictor is missing, a few more options abound: monotonic regression, predictive mean matching, or maximum likelihood estimation—however monotonic missingness is not the usual scenario.

2.10 Recapitulation

In December 2015, PCORI convened a workgroup to discuss missing data and data quality for research using electronic medical records and claims data—identifying problems, highlighting current solutions to some of those problems, and suggesting areas for future research [132]. One identified need was for research to understand the effects of various amounts of missing data, whether the results would be significantly altered by the amount of missing data. Another need was to understand which covariates experience missingness, what the mechanism for missingness is (e.g. MAR), and whether simulations could be used to learn more about these covariates and the impact of missing data. Finally, the PCORI workgroup asserted one of the next steps would be to bring researchers who have experienced success in handling missing data in EMR studies with researchers who are new to EMR studies to help disseminate this knowledge.

3 METHODS

3.1 Specific Aims and Hypotheses

This study has been designed to examine the effects of missing data in studies that use electronic health record data capturing patient stays in the intensive care unit for ventilator-dependent respiratory failure (VDRF) to accomplish the following:

- 1) Ascertain the degree to which results may be biased at various percentages of missing data
- 2) Examine methods of dealing with missing data that are commonly available in statistical software packages used by Health Services Researchers

Therefore, the aims of this study are as follows:

AIM 1

To examine the impact of missing SOFA score data on ICU clinical outcomes studies among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data at the composite score level.

Hypothesis 1: All methods for handling missing data will result in more accurate parameter estimates for all outcomes studied than simple pairwise deletion.

Hypothesis 2: Multiple imputation (MI) methods will result in the most accurate parameter estimates, when compared with the three other methods for dealing with missing data.

AIM 2

To examine the impact of missing SOFA score data on ICU clinical outcomes studies among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data at the component item level.

Hypothesis 1: All methods for handling missing data will result in more accurate parameter estimates for all outcomes studied than simple pairwise deletion.

Hypothesis 2: Multiple imputation (MI) methods will result in the most accurate parameter estimates, when compared with the three other methods for dealing with missing data.

3.2 Data Source

The data used in this study were provided by the Medical University of South Carolina's Clinical Data Warehouse (CDW), which contains electronic medical record data. These data include patient demographics, procedures, diagnoses, encounters, laboratory results, as well as medications ordered or administered during an inpatient admission [133]. The Medical University of South Carolina's CDW provides access to MUSC investigators with Institutional Review

Board approved studies for the purposes of retrospective research and patient recruitment [134]. The MUSC CDW has records collected since 1993 on 2.1 million patients, comprising 1.5 million inpatient admissions, 26.6 million outpatient encounters, 228 million laboratory results, and 5.5 million procedures as of October 2017 [135, 136]. These data were made available through funds from the South Carolina Clinical & Translational Research (SCTR) Institute, with an academic home at the Medical University of South Carolina, and a Duke Endowment Foundation Healthcare Division grant. These deidentified data were examined by the Institutional Review Board at the Medical University of South Carolina and deemed as non-human research.

The data used for this study contains demographics (age, sex, race), height, weight, primary payer, length of stay in the ICU, components of SOFA score upon admission to the ICU, duration of mechanical ventilation, diagnosis codes, procedure codes, discharge disposition, and total charges. Other pertinent clinical data were also retrieved, including Richmond Agitation-Sedation Scale (RASS) scores, Confusion Assessment Method for the ICU (CAM-ICU) scores, and results of Spontaneous Breathing Trials (SBT).

3.3 Study Population

The study population is composed of adults, aged 18 years or older on the date of admission, who were admitted to one of the ICUs at the Medical University of South Carolina from January 1, 2015 through October 31, 2017 and placed on a ventilator due to respiratory failure. Respiratory failure was defined as being indicated with an ICD-9 procedure code of 96.70, 96.71, 96.72; or an ICD-10 procedure code of 5A1935Z, 5A1945Z, 5A1955Z. The descriptions of these ICD procedure codes are listed below in Table 4.

Table 4 ICD-9 and ICD-10 procedure codes for study inclusion

<i>ICD Version</i>	<i>Procedure Code</i>	<i>Description</i>
9	96.70	Continuous mechanical ventilation of an unspecified duration
	96.71	Continuous mechanical ventilation of ≥ 96 consecutive hours
	96.72	Continuous mechanical ventilation of < 96 consecutive hours
10	5A1935Z	Respiratory ventilation, < 24 consecutive hours
	5A1945Z	Respiratory ventilation, 24-96 consecutive hours
	5A1955Z	Respiratory ventilation, > 96 consecutive hours

3.4 Statistical Software and Data Management

The data used for this dissertation was provided by the MUSC CDW team in comma separated value (CSV) format. The data were then imported into SAS software format, using SAS[®] software, version 9.4 for Windows (SAS Institute Inc., Cary, NC, USA). All simulations and analyses were performed using SAS/STAT[®] version 14.3 for the Microsoft Windows operating system.

3.5 Methods for Multiple Item Instruments

Instruments that have multiple component items present a decision point for the researcher, as one can handle these at the item (or component) level or the composite level. For instance, if a patient has values as shown below (Table 5), 4 out of 6 of the item level scores can be calculated.

Table 5 Example calculations of the SOFA score

<i>System</i>	<i>Value</i>	<i>Item Score</i>
Respiration PaO ₂ /FiO ₂ , mmHg	280	2
Coagulation Platelets x 10 ³ /mm ³	130	1
Hepatic Bilirubin, mg/dl	.	.
Cardiovascular MAP, mm Hg	68	2
Central Nervous System Glasgow Coma Scale	.	.
Renal Creatinine, mg/dl	1.1	0
SOFA Score		?

Unfortunately, the SOFA score itself—a summation of the 6 component scores—cannot be calculated in the case of one or more missing component scores. In the case of one or more missing component scores, the most one can say is that the SOFA score is at a minimum the sum of the observed component scores (i.e. $2 + 1 + 2 + 0 = 5$), and at most the observed component scores plus the maximum scores available for the missing component items (i.e. $2 + 1 + 2 + 0 + 4 + 4 = 14$). To further highlight the decisions available to the researcher, single imputation will be used for illustration.

In the case of single imputation for the above hypothetical scenario, the researcher can do single imputation for each component item missing, or for the composite SOFA score. However, this is complicated in the approach for the component items, as two options present themselves. The first option is to impute the missing values themselves (i.e. bilirubin level in *mg/dl*, and Glasgow Coma Scale score), then calculate the component scores—allowing the composite SOFA score to be calculated. The second option is to impute the component SOFA score—

allowing the composite SOFA score to then be calculated. The importance of this decision has been shown in the literature.

At least several studies have examined handling multiple item instruments at the composite versus component level. One study examined the Pain Coping Inventory, a 12-item instrument, using multiple imputation at both the composite and component levels [114]. This study found that when the percentage of missing data exceeded 25%, multiple imputation at the component level outperformed both mean imputation and multiple imputation at the composite level.

3.6 Aim 1 – Univariate Missingness (SOFA Score, Composite Level)

Aim 1 of this dissertation is to examine the impact of missing SOFA score data on ICU clinical outcomes studies among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data at the composite score level. The effects of various methods for handling missing data, described in subsequent sections, will be considered for their impact on the significance of three outcomes—both magnitude and direction—for three common outcomes that use SOFA score data as a risk adjuster. Essentially, using a dataset with complete SOFA scores, various percentages of missingness will be imposed so that we can compare the parameter estimates from various missing data techniques to those of the estimates from the full dataset.

3.7 Aim 2 – Multivariate Missingness (SOFA Score, Item Level)

Aim 2 of this dissertation is to examine the impact of missing SOFA score data on ICU clinical outcomes studies among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data at the component item level. The effects of various methods for handling missing data, described in subsequent sections, will be considered for their impact on the significance of three outcomes—

both magnitude and direction—for three common outcomes that use SOFA score data as a risk adjuster. Essentially, component SOFA score items will be deleted at various percentages of missingness to compare the parameter estimates from various missing data techniques to those of the estimates from the full dataset.

3.8 Simulation Process & Outcomes Analysis

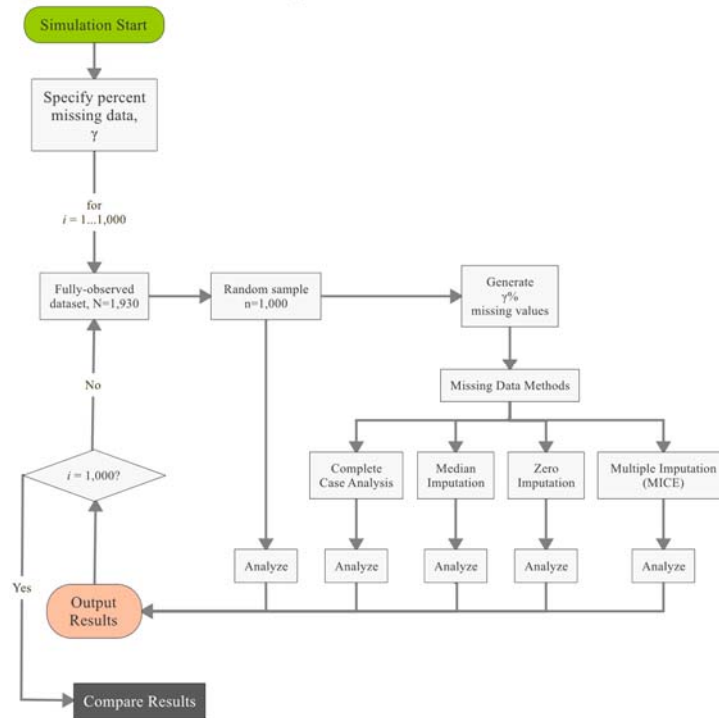
This study is designed as a simulation study to compare four statistical missing data methods to a known truth—the full dataset, to assess the performance of each. To aid in this endeavor, the published guide on conducting simulation studies in medical statistics by Burton et al. (2006) served as a guide for the simulation methods [137].

3.8.1 Simulation Algorithm

The simulation is represented in Figure 3.1 below. At the start of the simulation a complete dataset is provided; all component values of the SOFA score, composite SOFA score, potential covariates, and outcomes are fully-observed.

The two simulation parameters are specified. The first simulation parameter, S , denotes the number of simulation runs. Typically, S will be 1,000 runs—each looping through the simulation steps. The second simulation parameter, γ , denotes the percentage of observations within the dataset that will contain one or more missing values.

Once the simulation parameters have been specified and the fully-observed dataset chosen, the simulation loop proceeds as follows. From the complete dataset missing data will be generated by choosing γ percent of the observations to have a missing SOFA score value, and these values set to missing in the dataset. Then for each of the four missing data methods chosen, the missing data method will be applied to the dataset. Then it will be analyzed for the three outcomes chosen. For each analysis key results will be output to a table for later comparison. This simulation loop will run from the beginning, until S simulation loops have been completed.

Figure 3.1 Simulation algorithm

3.8.2 Simulation Parameters

There are two simulation parameters— S and γ . The first simulation parameter, S (not shown in diagram), denotes the number of simulation runs. Typically, S will be 1,000 runs—each looping through the simulation steps shown above in Figure 3.1. This creates S number of independent datasets from which the simulation can proceed.

The second simulation parameter, γ , denotes the percentage of observations within the dataset that will contain one or more missing values. For this study, a range of percent missing data will be studied to help understand the behavior of these data at various percentages of missingness. Studying a broad range of missing data percentages will further help Health Services Researchers to better understand how bias varies in these data based on the percent of missing data. Further, recommendations within the literature on the tolerable percentage of missing data

varies. Two guidelines state that 5% missingness is where one needs to worry about bias [12, 138], whereas another states 10% [139]. These guidelines are sufficiently broad, and varying, so further consideration in different types of data is warranted. To illustrate, while 5% missingness of the SOFA score may be reasonable within a large ICU dataset studying mortality, 5% missingness of race is likely excessive in disparities studies. Therefore, the tolerable percent missing would vary based on the missing predictor's strength of association with the outcome as well as the research question being asked.

3.8.2.1 Missing Data Mechanism & Generation of Missing Data

Data can rarely be considered to be missing completely at random (MCAR) due to the strict nature of these data—meaning it is akin to taking a random sample of the full data, dependent only on a stochastic process. Whereas one can never be fully certain whether data are missing at random (MAR) or missing not at random (MNAR). Even if one can model with good accuracy the missing data mechanism under MAR, there exists the possibility that data may also be missing due to a latent process—making those data MNAR. This was examined by Geert Molenberghs et al. (2008), when they demonstrated that any MAR missing data mechanism model has a corresponding MNAR model with equal fit, rendering empirical distinction between the two missing data mechanisms impossible [140]. Therefore, generation of missing data shall be conducted under the MAR and MNAR missing data mechanisms.

The MAR mechanism will be modeled using demographic missingness, similar to what has been used in other studies [102]. Using the original ICU dataset, which includes both missing and fully-observed data, we will model how demographic variables contribute toward missingness. If the existing data's missing data process were fully-MAR, then this would mimic the MAR missing data process. The method of accomplishing this is as follows.

If a subject's ICU record contains a missing SOFA score element, the record will be denoted as missing ($SOFA_{miss} = 1$), otherwise if all SOFA score elements are present then this will be denoted as ($SOFA_{miss} = 0$). A multivariable logistic regression model will then be fit to ascertain the estimated probability that a subject's SOFA score is missing, given their observed demographic and clinical factors—such as age, sex, race, and primary payor. Using the notation of Hosmer, Lemeshow, and Sturdivant (2013) [141], if there are p independent predictor variables, the vector of predictors is represented as $x' = (x_1, x_2, \dots, x_p)$ and the conditional probability that the SOFA score is missing is denoted as follows:

$$\Pr(SOFA_{miss} = 1|x') = \pi(x')$$

From the multiple logistic regression model, we obtain parameter estimates for each of the p predictors, yielding the logit transformation, $g(x)$, shown below:

$$g(x') = \ln \left[\frac{\pi(x')}{1 - \pi(x')} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Finally, assigning the empirically-derived probability of having a missing SOFA score to each record will be defined as follows:

$$\Pr(SOFA_{miss}) = \pi(x') = \frac{e^{g(x')}}{1 + e^{g(x')}}$$

These values will not change across simulation iterations. However, for each run through the simulation (depicted in Figure 3.1), random variation will be introduced to these probabilities. To accomplish this, a random component R will be added to $\Pr(SOFA_{miss})$. This random component will consist of a random draw from the standard uniform distribution, where $R \in [0,1]$, by using the SAS function $RAND('UNIFORM',0,1)$. The fully-observed dataset will then be sorted in descending order by its probability of missing, with the first $\gamma\%$ of observations having their SOFA score (or 1 or more SOFA score components) set to missing.

The MNAR mechanism will be modeled using three basic strategies. The first strategy will be to impose missingness more commonly in the lower SOFA score categories, in the left side of the observed SOFA distribution (hereinafter referred to as *MNAR Left*). This means that SOFA scores closer to zero would have a higher likelihood of being deleted. This strategy aligns well with the SOFA score creators' guidelines to impute a zero when the score (or one of the sub-scores) is missing.

The second strategy will impose missingness more in the median SOFA score categories, in the center of the observed SOFA distribution (hereinafter referred to as *MNAR Mid*). Finally, the third strategy will be to impose missingness in the higher SOFA score categories, in the right side of the observed SOFA distribution (hereinafter referred to as *MNAR Right*). These will be accomplished using the methods outlined by Jaap Brand et al. (1993) using SAS/STAT software [142], yielding $\gamma\%$ of the fully-observed dataset having missing SOFA score observations.

3.8.2.2 Assignment of Missing Data Patterns

The assignment of missingness patterns proceeded as follows. First for each observation within the fully-observed dataset, a missing data pattern was assigned to each observation according to the frequencies with which these patterns occurred within the dataset extracted from electronic health record. Then, observations were chosen to be selected to have missing data according to the respective missing data mechanism and percentage of missingness as prescribed by the parameters within the simulation. If an observation was selected to be missing, the missing data pattern that was previously assigned was then applied.

3.8.2.3 Simulation Runs & Percent Missingness

As mentioned previously, two simulation parameters are specified: the number of simulation runs (S), and percentage of observations within the dataset that will contain one or more missing values (γ). It is common within simulation studies that the number of simulation

runs be set to a large number ranging from 200-1,000 [142]. Of the simulation studies examined for this research, 1,000 simulation runs seemed to be the most common [102, 143-145], followed by 500 simulation runs [114, 142]. Therefore, in this study the number of simulation runs, S , will be set to 1,000—each looping through the simulation steps shown in Figure 3.1.

The second simulation parameter that will be specified, γ , denotes the percentage of observations within the dataset that will contain one or more missing values. In this study, any observation that has one or more items of the SOFA score missing contributes toward the number of missing observations.

The most common guideline has asserted that >5% missingness is the level where a researcher needs to worry about biasing the results of a study [12]. While this particular percentage of missingness merits further investigation in these data, it would be wise to go much higher than this amount. Further, sufficiently small steps to aid in decision making should be taken between the percentage of missingness levels. Therefore, in this research project percentages of missingness from 0% to 40% at 10% increments were investigated; we investigated $\gamma=0\%$, 10%, 20%, 30%, and 40% missingness. The baseline of 0% missingness—the fully-observed dataset—was used as the referent group from which *truth* was derived, and all comparisons in this study were made.

3.8.3 Missing Data Methods

In this study we investigated four methods for handling missing data. These methods include the most common method used in health services research, pairwise deletion (complete case analysis), two deterministic imputation techniques (median imputation and imputation per SOFA guidelines), and multiple imputation. Maximum likelihood estimation was not be considered due to the assumption of multivariate normality, and the limitation of only being able to model continuous outcomes using standard software—such as SAS, SPSS, and Stata; one

would have to use specialized software such as Mplus for such applications of generalized linear models [146]. As all three outcomes compared in this analysis required generalized linear models (see Section 3.8.4), MLE was not considered for comparison in this study.

Therefore, the effects of these four methods for handling missing data were considered for their impact on the significance of three outcomes—both magnitude and direction—that use SOFA score data as a risk adjuster. These methods will be operationalized at the composite SOFA score level (in support of Aim 1), and at the component SOFA score level (in support of Aim 2). These missing data methods will be briefly explored below.

3.8.3.1 Method 1: Complete Case Analysis

The first method for handling missing data that that was explored in this study is complete case analysis (CCA), whereby only those cases for which data exists on all outcomes and potential explanatory variables are retained in the analysis. The second method is available case analysis (ACA), also known as pairwise deletion, which is the default method of most statistical software. With this method any observation that has a missing outcome or explanatory variable is excluded from the analysis. This leads to problems of changing samples, as during model fitting as potential covariates are added and removed the sample size will change. Further, secondary analyses will also have different sample sizes—and essentially different samples—than the primary analysis. Clearly, such a strategy is problematic as will be discussed next.

The first problem with CCA and ACA methods is that reduced statistical power is achieved due to a smaller sample size. While the magnitude of this problem varies based on the total sample size and the percent of missing data, it is still a problem. The second problem with CCA and ACA is that unless the data that are missing are missing completely at random (MCAR), the results will likely be biased [109]. If it is completely a chance occurrence the data are missing— or MCAR—removing the case will not bias the results using CCA [109]. However,

CCA and ACA will result in a loss of precision in the confidence interval. Therefore, dealing with the missing data, rather than ignoring it, is warranted. However, in order to show the magnitude of bias at varying percentages of missing data, as well as mechanisms of missingness, it is imperative that we explore this as a missing data method herein.

For this study any observation where the composite SOFA score is missing (in the case of Aim 1), or any one component of the SOFA score is missing (in the case of Aim 2) will be deleted. Then three outcomes will be analyzed, as described in Section 3.8.4.

3.8.3.2 Method 2: Median Imputation

The second method for handling missing data that was explored in this study is median imputation. Median imputation is a deterministic imputation technique that creates one complete dataset by filling in the missing values to allow all the data to be used. Essentially, the researcher simply imputes the median value of the missing item and proceeds with analysis as if no data were previously missing.

Granted, this method has been rejected by Rubin as being unacceptable for research [113], and has been repeatedly shown in studies to introduce unacceptable bias and over-precise confidence intervals [114]. However, it is important that this method be demonstrated within this current research for the same reason that pairwise deletion will be used—this method is still being used in scientific studies.

For this study the median composite SOFA score (in the case of Aim 1), or the median component SOFA score (in the case of Aim 2) will be imputed in cases of missing values. Then three outcomes will be analyzed, as described in Section 3.8.4.

3.8.3.3 Method 3: Imputation per SOFA Guidelines (Zero)

The third method for handling missing data that was explored in this study was to impute a zero for each missing composite SOFA score (in the case of Aim 1), or a zero for each

component item missing (in the case of Aim 2). This methodology of assuming the score is zero, meaning there is no organ dysfunction, is in line with the SOFA score guidelines, outlined in the 2016 Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) consensus paper. According to the Sepsis-3 consensus paper, the baseline SOFA score—which is calculated upon admission to a critical care unit—should be assumed to be zero, unless the patient has a known organ dysfunction [65]. In the quantitative study which informed the Sepsis-3 task force, single imputation of a normal value (zero) was used [147], with the assertion made that this usage “mirrors how clinicians would use the score at the bedside” (p. 764). While this observation has merit, it is worth noting that a treating clinician would have access to information—the full clinical picture—to which a later researcher would not have access.

Similar to median imputation, imputing a zero for data that are missing is a deterministic imputation technique that creates one complete dataset by filling in the missing values with zeroes to allow all the data to be used. For this missing data method, any observation where the composite SOFA score is missing (in the case of Aim 1), a zero will be imputed. For cases where any component of the SOFA score is missing (in the case of Aim 2), a zero will be imputed for missing component scores. Then three outcomes will be analyzed, as described in Section 3.8.4.

3.8.3.4 Method 4: Multiple Imputation

The fourth method for handling missing data that was explored in this study is multiple imputation. As discussed in-depth in Chapter 2, multiple imputation creates multiple datasets—imputing values with a random component added, which represent the variation that we might expect when sampling from a population. These datasets are then analyzed using normal analysis techniques. Finally, the point estimates and standard errors are combined using standard combining rules. Multiple imputation is used when the MCAR or MAR missing data mechanisms are likely, but have been used with MNAR.

The goal of MI is not to make up data, but rather to allow all the data that are present to be used in analyses to achieve valid statistic inference, not perfect point prediction [113]. Essentially, MI fills in the values that are missing with plausible values, to allow all of the existing data to be used—rather than discarded.

There are two common methods of MI for multivariate missing data that could be considered for this study, multivariate imputation through chained equations (MICE) and Markov-Chain Monte Carlo (MCMC) multiple imputation. Multivariate imputation through chained equations is a type of MI whereby various regression-based techniques—typically generalized linear models, such as linear or logistic regression—can be used for each type of missing variable in the data. In other words, a logistic regression model can be used to help predict missing binary values; an ordinary least-squares (linear) regression model can be used to help predict missing continuous variables.

The second type of MI is Markov Chain Monte Carlo MI (MCMC), which assumes multivariate normality of all continuous variables [12]. The MCMC method specifies a single joint distribution for all variables with continuous data, which would be difficult to achieve. Particularly in attempting to impute missing SOFA scores, as the components thereof are count data rather than continuous, which would be required to specify a single joint distribution.

Multivariate imputation through chained equations has advantages in health services research studies due to the wide range in the types of variables for which we might control in a regression analysis, allowing separate imputation models to be used for each variable with missing data. Typically, continuous, ordinal, categorical, and binary variables will all be used in the same analysis as covariates—demonstrating the appeal of using MICE. For the foregoing reasons, MICE will be the MI method explored herein.

The number of imputations, denoted by m , is also of concern. The recommendation by Rubin in 1987 was that between 2 and 10 imputations are sufficient [98], which he reemphasized a decade later [113] asserting $m = 3$ or 5 is often sufficient.

The ground for Rubin's assertion comes from his earlier work, which states that the relative efficiency (in standard deviations) of m imputations and γ percent missingness, compared to $m=\infty$ is estimated as $\frac{1}{\sqrt{1+(\gamma/m)}}$ [98]. As an example, with 30% missing data, and 25 imputations

performed ($m=25$), the relative efficiency is 99.4%. This is a small increase over simply performing 10 imputations ($m=10$) at $\gamma=30\%$, which has a relative efficiency of 98.5%.

Nonetheless, some of the more recent literature have demonstrated *imputation variability* with smaller numbers of imputations, which coupled with the modest increase in processing time for doing larger numbers of imputations make the case for doing more imputations than previously recommended [148]. This thinking has been adopted by major statistical software packages, such as SAS—which now defaults to $m=25$. For the MICE method, the default number of multiply-imputed datasets in SAS ($m=25$) will be used. Finally, the m number of datasets—specifically the m number of parameter estimate vectors and standard error vectors—will be combined using Rubin's rules [98].

3.8.4 Analysis of Outcomes

There are three outcomes to be analyzed to compare the performance of the four methods for handling missing data—death, total hospital charges, and the length of stay within the ICU. Comparisons of the missing data methods were made using these three outcomes as indicators of missing data method performance. The various missing data methods were compared to results from the full dataset for these three outcomes. These three outcomes will be discussed in greater depth in the sections that follow.

For all of these analyses, the fully-observed model were fit first—prior to simulation runs—to find parsimonious models that best predict each of the three outcomes. Then the chosen model for each of the outcomes were used in the simulation runs without further model fitting. Details of model fitting for each outcome, along with the importance of each outcome, will be discussed next.

3.8.4.1 Outcome 1: Death

The outcome of in-hospital death is important, and often studied when using SOFA score as an instrument for baseline severity assessment. While the authors of the SOFA score emphatically asserted that it was designed as a tool for description, not prediction, [64] its usage has changed over time as it has demonstrated to be a good prognostic tool among ICU patients. As discussed in *Section 2.5.3*, the SOFA score is widely-used as a prognostic tool to predict in-hospital (and in-ICU) death. As such, investigating the impact of missing SOFA score data on the clinical outcome of in-hospital death among ICU patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data is important.

This model will be fit with a multiple logistic regression model. This analysis will be performed using PROC LOGISTIC with SAS/STAT. The first model will be fit on the full-observed dataset, then the same model will be used in all of the simulations without subsequent model fitting.

Multicollinearity between covariates will be evaluated using Spearman's rank correlation coefficients ($\hat{\rho}_s$), along with variance inflation factors (VIF) during modeling. Model fitting will proceed as described by Hosmer, Lemeshow, and Sturdivant (2013), using manual backwards selection, using the smallest Akaike Information Criterion (AIC) or Schwarz Criterion (SC), and likelihood ratio tests, along with individual covariate significance (p-values) [141]. Clinically-

relevant variables will be used to determine which covariates will initially be included in the model to control for differences in comparison groups and potential confounding. Overall model fit will be assessed using the Hosmer-Lemeshow goodness of fit test. Model assumptions of linearity of the logit for continuous predictor variables will be assessed using graphical methods (LOWESS scatterplot). Statistical significance for this outcome was determined *a priori* to be at the $\alpha=0.05$ level.

3.8.4.2 Outcome 2: Total Charges

Studies that examine total charges are useful in economic evaluations of treatment, such as cost effectiveness analysis studies. However, without severity adjustment such evaluations will have large confidence intervals in point estimates of cost savings. Herein shall be examined one study that did not adjust for severity within the ICU, and another study that did.

The first study is one that did not adjust for patient severity within the ICU using a validated severity of illness scoring system, such as the SOFA score. is a random-effects meta-analysis of 12 studies that examined early versus late tracheostomies among ventilated ICU patients, showed the average ICU cost difference was \$4,316 (95% CI: \$403-8,229), favoring early tracheostomies [149]. However, in this study severity adjustment was not made—which might have shrunk those confidence intervals, to give better estimates of the cost difference between early versus late tracheostomy among ventilated ICU patients.

The second study is one that did adjust for patient severity within the ICU using a validated severity of illness scoring system. In this study the authors investigated, via simulation, how to optimize telemedicine delivery to ICUs (Tele-ICU) based on patient severity of illness—as measured by the Acute Physiology and Chronic Health Evaluation IV (APACHE-IV) score [150]. The authors wanted to know at what severity of illness—typically associated with poorer outcomes and higher costs of care—would tele-ICU prove to be cost effective. The study found

that using Tele-ICU among the 30-40% of highest risk patients demonstrated optimal cost effectiveness, with an incremental cost effectiveness ratio (ICER) of \$25,392 per quality-adjusted life year (QALY). This study demonstrated the use of telemedicine among ICU patients with the 30-40% of highest severity could be cost effective, which could inform an intervention study.

In the first study mentioned, cost effectiveness of an intervention (early tracheostomy) was studied—without adjusting for patient severity, which likely led to imprecision in cost savings. The second study of cost effectiveness of tele-ICU was investigated, using patient severity as a criterion for tele-monitoring; the optimal patient severity to demonstrate cost effectiveness was found, which can aid decision-makers and clinicians.

Cost effectiveness studies will continue to be performed, as balancing the allocation of resources among treatment alternatives, and patients, remains a priority in health services research. In all such economic studies, severity of illness (along with other potential confounders) should be adjusted for in analyses. As such, investigating the impact of missing SOFA score data on the financial outcome of total hospital charges among ICU patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data is important.

Total charges measure the acute episode of care, both professional and facility charges. These charges do not include follow-up care, including post-acute, specialty referrals, nor any therapy. To quantify the effect of SOFA scores (e.g. a patient's severity of illness) on the total charges—which is a measure of intensity of care—generalized linear modeling will be used.

This model will be fit with a gamma-distributed generalized linear model with a log-transformed link function. In a gamma distribution, the standard deviation of the outcome is proportional to the mean ($\sigma \propto \mu$) [151]. Further, the generalized linear model has been shown to

work well in healthcare cost studies, where costs are heavily right-skewed, as long as the log-transformed variable does not have excessive heteroscedasticity [152].

This analysis will be performed using PROC GENMOD with SAS/STAT. The first model will be fit on the fully-observed dataset, then the same model will be used in all of the simulations without subsequent model fitting. Multicollinearity between covariates will be evaluated using Spearman's rank correlation coefficients ($\hat{\rho}_s$), along with variance inflation factors (VIF) during modeling. Initial modeling will proceed with manual backwards selection, using the smallest Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) values, and individual covariate significance (p-values). Clinically-relevant variables will be used to determine which covariates will initially be included in the model to control for differences in comparison groups and potential confounding. Statistical significance for this outcome was determined *a priori* to be at the $\alpha=0.05$ level.

3.8.4.3 Outcome 3: ICU Length of Stay

The outcome of ICU length of stay is clinically-important. One large benchmark study of all ICU admissions at 271 ICUs in the United States in 2008 found that for each day in the ICU, a patient will spend 1.5 days in a non-ICU bed [153]. With the typical cost per day of an ICU stay being \$3,518 in 2005, and the typical non-ICU stay at \$1,153 [2], reductions in ICU length of stay can reap large reductions in healthcare expenditures. As such, investigating the impact of missing SOFA score data on the clinical outcome of length of stay within the ICU among patients with ventilator-dependent respiratory failure at various percentages of missingness, along with various statistical techniques for handling missing data is important.

This model will be fit with either a negative binomial- or Poisson-distributed generalized linear model, depending upon model fit. Typically, for count data a Poisson-distributed model will be appropriate when the variance of the outcome equals the mean, whereas a negative

binomial-distributed model will be appropriate when the variance of the outcome is a quadratic of the mean [151]. Selection of appropriate model—negative binomial or Poisson—will be assessed by comparing the model’s deviance per degree of freedom, $\frac{Deviance}{d.f.}$, with the chosen model exhibiting a value closest to unity (1.0). This analysis will be performed using PROC GENMOD with SAS/STAT. The first model will be fit on the fully-observed dataset, then the same model will be used in all of the simulations without subsequent model fitting.

Multicollinearity between covariates will be evaluated using Spearman’s rank correlation coefficients ($\hat{\rho}_s$), along with variance inflation factors (VIF) during modeling. Initial modeling will proceed with manual backwards selection, using the smallest Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) values, and individual covariate significance (p-values). Clinically-relevant variables will be used to determine which covariates will be initially added to the model to control for differences in comparison groups and potential confounding. Statistical significance for this outcome was determined *a priori* to be at the $\alpha=0.05$ level.

3.8.5 Output of Results from Simulations

For each analysis, parameter estimates ($\hat{\beta}$), standard errors (SE), sample size used in the analysis, and the calculated variance of the parameter estimate ($\widehat{\sigma}^2 = (SE/\sqrt{n})^2$) will be output to a table along with identifiers of the simulation run (i, γ , identifier for the outcome being modeled [e.g. death], and the missing data method used) for later comparison. An example of the output table is below, in Table 3.3.

Table 6 Simulation output table (example)

<i>Row ID</i>	<i>i</i>	<i>γ</i>	<i>Missing Data Method</i>	<i>Modeled Outcome</i>	<i>β Estimate</i>	<i>Standard Error</i>	<i>n</i>	<i>Variance</i>
1	0	0	Full Dataset	Death	0.406	0.179	1000	32.005
2	1	20	Complete Case	Death	0.559	0.135	800	14.558
3	1	20	Multiple Imputation	Death	0.414	0.162	1000	26.374
4	1	20	Guidelines (Zero)	Death	0.429	0.198	1000	39.363
5	1	20	Median	Death	0.465	0.206	1000	42.601

As one can see from the above Table, there are 5 rows, where column i represents the number of the iteration through the simulation loop. Where $i=0$, these are the initial analyses on the fully-observed dataset, against which comparisons will be made. Column γ denotes the percent missing data generated in that run. Therefore, with 4 missing data methods chosen for comparison in the simulation loop, each simulation loop iteration will yield 4 rows (as shown above) for each outcome being modeled, which will be output to the results table for each outcome being analyzed; for a simulation run where $S=1,000$ with 4 missing data methods and 3 outcomes being modeled, 12,000 rows ($1,000 * 4 * 3$) will be output to a table. These statistics will be used to compute the test statistics described in the next section, allowing for comparison of methods.

3.8.6 Assessment of Simulation

The simulation runs were compared using summary statistics of the three test statistics, yielding the properties of the performance of each of the missing data methods examined. These three statistics are described below.

1. **Relative Bias** is calculated as $\frac{\hat{\beta}_i - \beta}{\beta}$.

Relative bias will be calculated for each simulation run and missing data method. With this, $\hat{\beta}_i$ represents the parameter estimate for the SOFA score for simulation run i for each of the $i = 1, \dots, S$ simulation runs in the generalized linear model, and β represents the

population parameter of the SOFA score from the fully-observed dataset. Ideally, this number will be 0%—meaning no bias exists in the missing data technique incorporated. Means and 95% confidence intervals of the relative bias for each missing data technique at each percentage of missingness will be calculated, allowing the observation of the magnitude and direction of bias that a missing data technique introduces.

2. **Efficiency** is calculated as $\frac{\hat{\sigma}_i^2}{\sigma^2}$.

Efficiency will be calculated for each simulation run and missing data method. Efficiency is a simple ratio of the variance of the parameter estimate of the SOFA score $\hat{\beta}_i$, denoted as $\hat{\sigma}_i^2$, compared to the variance of the parameter estimate for the SOFA score in the fully-observed dataset β , denoted as σ^2 . Means and 95% confidence intervals of the efficiency for each missing data technique at each percentage of missingness will be calculated.

3. **Coverage Probability** is the proportion of simulation runs for each missing data method where the parameter estimate for the SOFA score in the fully-observed dataset β , is contained within the confidence interval for the estimated coefficient $\hat{\beta}_i$. Since $\alpha=0.05$ has already been set for this study, we desire a $1-\alpha = 95\%$ or greater coverage probability. In general, for this alpha level 95% coverage is considered to be ideal, whereas coverage of less than 95% is indicative of higher than expected Type-I error rate [137].

3.8.7 Monitoring of Simulation Process

For each simulation run one log was output to monitor the status of each simulation, along with five tables. The simulation log contained all notes, warnings, and errors for each simulation run, along with the times to accomplish each of the critical steps.

The first table created was the fully-observed dataset's parameter estimates for all three outcomes. These parameter estimates were used as the truth against which each simulation method's parameter estimates would be compared.

The second table created contained the convergence status of each of the analyses was conducted. This table was later analyzed to ensure all models converged.

The third table created contained all seeds used in the multiple imputation process. All seeds for the multiple imputation processes were generated using a hybrid 1998/2002 32-bit Mersenne twister pseudorandom number generation algorithm in SAS, exhibiting good statistical qualities in mimicking a stochastic process [154]. This pseudorandom number generator has a period of $2^{19,937}-1$, which is the number of calls to the pseudorandom number generator one would have to make in order to have a repeated sequence of values [155]. In spite of the infinitesimally small probability of seed repetition across the simulation runs, the seeds generated in each simulation loop were saved to a table for later evaluation of repetition. As expected, the 40,000 random number seeds generated were all unique integers, with no duplicates found across the simulation runs—ensuring good statistical independence of each multiple imputation run.

The fourth table created contained the parameters from the multiple imputation process, along with the variance of the parameter estimates used for multiple imputation.

Finally, and most critically, the fifth table created contained the parameter estimates from each of the three analyses across the four missing data methods. The information within this table was used to calculate the summary statistics used in this study—coverage, relative bias, efficiency, and root mean square error in fulfillment of Aim 1 and Aim 2 of this study.

For each simulation run, approximately four minutes were required to build the 1,000 datasets and impose missingness in accordance with the simulation parameters (i.e. missing data mechanism, and percent missingness). Each of the first three methods for handling missing data

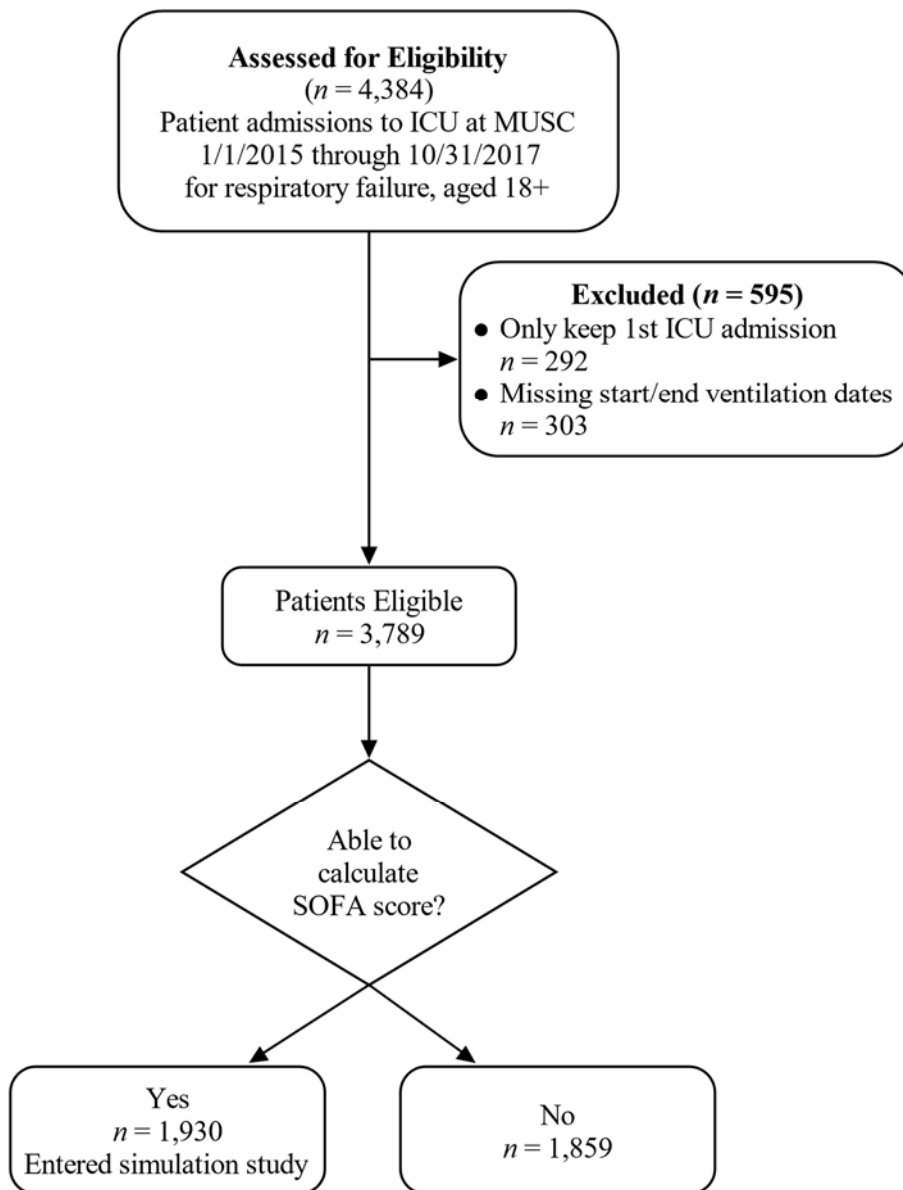
took approximately 30 seconds to run for the 1,000 datasets—including both handling the missing data and performing the analyses.

However, for the multiple imputation process, the multiple imputation of the 1,000 datasets took approximately four hours to run. This is predominantly due to the scaling of the analyses; for each dataset, the multiple imputation process is repeated 25 times. Therefore, for the 1,000 datasets in a single simulation, multiple imputation occurred 25,000 times. Following the multiple imputation process, these 25,000 datasets were analyzed separately for each of the three outcomes, with the results of each multiple imputation process being combined using Rubin's Rules—all of which occurs in one step in SAS using PROC MIANALYZE. Each analysis took approximately five minutes to run. Therefore, the total time to run one full simulation was approximately five hours. Given that this study examined two aims, each having 4 missing data mechanisms (MAR, MNAR Left, MNAR Mid, MNAR Right) and 4 levels levels of missing data ($\gamma = 10\%$, 20%, 30%, 40%), a total of 32 simulations were ran, consuming approximately 160 hours of processing time.

4 RESULTS

4.1 Data Used in Dissertation

From the MUSC Enterprise Data Warehouse records for 4,384 patient admissions to the various ICUs at MUSC from the period of January 1, 2015 through October 31, 2017 for ventilator-dependent respiratory failure among patients 18 or older were extracted. Of these patients, 292 had multiple ICU admissions — therefore, only the first admission to an ICU at MUSC were used. An additional 303 of these admissions were missing either the start or end of ventilation dates resulting in exclusion from this study due to the inability to calculate total time on a ventilator. This resulted in 3,789 patients being eligible for the study. Of these patients, 1,930 (50.9%) had sufficient data to calculate a full SOFA score; 1,859 of these patients (49.1%) were missing one or more data elements required to calculate the SOFA score. The former of these two groups will hereinafter be referred to as the *fully-observed cohort*, whereas the latter of these two groups will be referred to as the *partially-observed cohort*. This process of building the cohort from which this simulation study would pull—the fully-observed cohort—is depicted below in Figure 4.1.

Figure 4.1 Data flow diagram

4.1.1 Descriptive Characteristics

A total of 3,789 unique patient admissions to the ICU were extracted from the Medical University of South Carolina's electronic health record during the time period of January 1, 2015 through October 31, 2017 and placed on a ventilator due to respiratory failure. Of these patient admissions, sufficient data elements were present within the data extract to calculate SOFA scores for 1,930 patient admissions, whereas insufficient data were present to calculate SOFA scores for 1,859 patient admissions—representing 49.1% missingness within the original data extract. The demographics and characteristics of these two groups of patients, grouped on whether or not the SOFA score was able to be calculated, are represented below in Table 7.

Table 7 Demographics and characteristics of patients in the original dataset

	<i>SOFA Score Present?</i>		<i>p-value</i> [†]
	<i>Yes</i> (<i>n</i> = 1,930)	<i>No</i> (<i>n</i> = 1,859)	
Age, years	56.6 ±17.1	55.6 ±17.7	0.0844
Male	1,126 (58.3)	1,099 (59.1)	0.6277
Race			0.1631
Black	794 (41.1)	735 (39.5)	
White	1,042 (54.0)	1,051 (56.5)	
Other/Unknown	94 (4.9)	73 (3.9)	
Insurance			0.9739
Commercial	598 (31.0)	577 (31.0)	
Medicare/Medicaid	1,078 (55.9)	1,042 (56.1)	
Other/Unknown	254 (13.1)	240 (12.9)	
Charlson Score	3.0 ±3.0	3.2 ±2.8	0.0919
Length of Stay ^a			
ICU	10.0 ±10.7	9.9 ±11.3	0.8137
Overall	17.1 ±21.6	17.3 ±19.4	0.7547
Total Charges	\$198,539 ±219,757	\$200,896 ±225,994	0.7449
Died	704 (36.5)	507 (27.3)	<0.0001
SOFA Score	8.8 ±4.1 8 [6] ^b		
SOFA Components ^c			
CNS	2 [0-4]		
Cardiovascular	1 [0-4]		
Coagulation	0 [0-4]		
Hepatic	0 [0-4]		
Renal	1 [0-4]		
Respiratory	3 [0-4]		

Note. All values are expressed as mean ±S.D., n (%), or as otherwise indicated.

[†] P-values were calculated using the Wilcoxon Mann-Whitney U test for continuous measures, and the χ^2 or Fisher's Exact tests for categorical measures (as appropriate). Statistically-significant comparisons at the $\alpha=0.05$ level are given in **bold**.

^a Expressed in days

^b median [Interquartile range]

^c median [Range]

Other/Unknown race is comprised of Asian, Hawaiian, Indian/Alaskan, and where this value was missing from the original dataset.

Comparison of the demographics and characteristics of these two groups of patients reveals very little differences between the two groups. These two groups did not differ significantly with respect to age, sex, race, insurance status, Charlson score, length of stay, nor total charges. However, these groups did differ in the overall death rate. The overall death rate was 36.5% among the patients for whom we were able to calculate a SOFA score, whereas this rate was 27.3% among those for whom we could not calculate a SOFA score ($p < .0001$).

As a higher SOFA score is indicative of greater patient acuity, this statistically-significant difference suggests that the SOFA scores among those patients who have a missing SOFA score in this data extract would have been on the lower end of the scale, meaning that these patients had less organ derangement and were therefore more healthy and less likely to die. Furthermore, this suggests that the missing data mechanism in these data may be dependent upon the SOFA score value, as those patients whose SOFA scores are missing have a lower rate of death, which would be evidence toward concluding the missing not at random (MNAR) missing data mechanism may be present. However, as with any missing data process, other factors may be influential or this statistically significant difference may be spurious.

4.1.2 Bivariate Analyses

Bivariate analyses examining the relationship between the SOFA score, components of the SOFA score, and all potential predictors was undertaken. Spearman's rank correlation coefficients ($\hat{\rho}_s$) were calculated, along with p-values. These measures are given in Table 24 of *Appendix C. Correlation table*, on page 154. Interpretation of the coefficients is according to Shi (2008; p. 371; [156]), as shown below in Table 8.

Table 8 Interpretation of Spearman's rank correlation coefficients

$ \hat{\rho}_s $	<i>Strength of correlation</i>
0.00 to 0.19	Little to none
0.20 to 0.39	Slight
0.40 to 0.59	Substantial
0.60 to 0.79	Strong
0.80 to 1.00	Very Strong

$|\hat{\rho}_s|$ are given, with sign indicating direction of relationship, where $\hat{\rho}_s \geq 0.01$ indicates a positively-correlated relationship and $\hat{\rho}_s \leq -0.01$ indicates a negatively-correlated relationship. This table is adapted from Shi (2008), page 371.

4.1.3 SOFA Scores

In this section, an examination of the possible missing data mechanism of the SOFA score, the missing data patterns observed for the components of the SOFA score, as well as the distribution of the SOFA score within the fully-observed dataset will be examined.

4.1.3.1 Missing Data Mechanism

A logistic regression model was fit to determine which covariates might predict missingness of the SOFA score in the original data. A binary indicator variable indicating whether or not the SOFA score was able to be calculated (*SOFA_missing*), served as the primary outcome variable for the logistic regression model, with all potential and relevant demographics and clinical outcomes measured in this study used as potential predictors of missingness of the SOFA score. These potential predictors tested were: discharge disposition, race, payor group, age group, sex, ICU length of stay, Charlson score, and the natural logarithm of total charges. For records with total charges equal to \$0.00 (n=2; one from each *SOFA_missing* group), these charges were changed to \$1.00. The minimum RASS score and the maximum CAM-ICU score during the first two days of ventilation were investigated for potential prediction ability. However, these

variables were missing in 5.0% and 36.6% of the cases respectively, prohibiting inclusion into the prediction model.

The final parsimonious model was selected using a backwards stepwise process, retaining only those predictors that had a statistically significant ability to predict a SOFA score to be missing (*SOFA_missing* = 1). The final model included only one variable, discharge disposition. Discharge disposition was divided into three groups: (1) died or sent to hospice, (2) sent home, and (3) sent to institutionalized care. The Hosmer-Lemeshow goodness of fit test showed excellent model fit, with a p-value of 1.0.

The logistic regression model to predict a missing SOFA score showed statistically-significant differences in missingness were found in the disposition groups. Patients who died or were sent to hospice care had lower odds of having a missing SOFA score when compared with patients who were discharged home (OR 0.571, 95% CI: 0.491-0.665). Similarly, patients who were discharged to institutionalized care had lower odds of having a missing SOFA score when compared with patients who were discharged home (OR 0.764, 95% CI: 0.651-0.897). The results of the final logistic regression model predicting a missing SOFA score are shown below in Table 9.

Table 9 Odds ratios and 95% confidence intervals for predicting a SOFA score being missing in the original data

<i>Effect</i>	<i>Odds Ratio</i>	<i>95% Confidence Interval</i>	<i>p-value</i>
Discharge Destination			
Died/Hospice	0.571	0.491-0.665	<0.0001
Institutionalized	0.764	0.651-0.897	0.0008
Home	†		

Note. Odds ratio estimates and Wald 95% confidence intervals given. Statistically-significant comparisons at the $\alpha=0.05$ level are given in **bold**.

† indicates the reference group.

Patients in the ICU who died, were sent to hospice, or were institutionalized are patients who, in general, would be expected to have poorer outcomes in comparison with those ICU patients who were sent home. This former group of patients who have poorer outcomes would, in general, have higher SOFA scores; the latter group of patients who are sent home would generally have better health outcomes and lower SOFA scores. Certainly, the group who experienced death as their discharged destination had a poor outcome, and likely higher rates of organ derangement, as indicated by a higher SOFA score. However, for those who were also sent to hospice or who were institutionalized following their ICU stay, their discharge destination would be demonstrative of a poorer outcome than a patient who was deemed well enough to go home by the Intensivist. Moreover, as a higher SOFA score positively correlates with greater patient acuity within the intensive care setting as well as poor outcomes following the ICU stay, it is probable that the SOFA scores for the patients whose scores we were unable to calculate as a result of one or more missing components would have, on average, lower SOFA scores than those whose SOFA scores we were able to calculate. Therefore, if this is the case, a missing not at random (MNAR) mechanism is

present within the data—where lower SOFA scores have a greater likelihood of being missing. Thus, special attention should be paid to results of the MNAR simulations at the low end of the SOFA score range (*MNAR Left*). However, this finding is given with caution as other studies have shown a single mechanism is unlikely to be the sole cause of missingness [102]. Moreover, any MAR missing data mechanism model has a corresponding MNAR model with equal fit, rendering empirical distinction between the two missing data mechanisms impossible [140].

4.1.3.2 Missing Data Patterns

An examination of the missing data patterns that were present within the data was conducted using PROC MI in SAS. This examination revealed 1,859 of the 3,789 ICU admissions (49.1%) had one or more items from the SOFA score that were missing (c.f. Table 10). The hepatic component of the SOFA score (measured by bilirubin) was the item most commonly missing; this finding matches that of another study [94]. The central nervous system component (measured by the Glasgow Coma Scale) was the second most common missing item. The cardiovascular component (measured by mean arterial pressure [MAP] and vasopressors) was the most infrequently missing; this differs from another study, which found platelets (coagulation SOFA component) and PaO₂/FiO₂ ratios (respiration SOFA component) to be the most infrequently missing items [94].

In total, there were 45 distinct missing data patterns across the six component items of the SOFA score, out of the $2^6 = 720$ possible patterns. The top 25 most common missing data patterns—which accounts for 92.8% of all observations with missing SOFA score data—are shown below in Table 11. Of the 45 missing data patterns, only

10 patterns exceeded more than 1% of the total missingness of the data. Therefore, these 10 missing data patterns along with their observed frequencies within the original dataset are used as the missing data patterns for this simulation study.

Table 10 Frequency of missing SOFA score components in original data

SOFA Component	# Missing	% Missing
Central Nervous System	753	19.9
Cardiovascular	113	3.0
Coagulation	215	5.7
Hepatic	1,040	27.5
Renal	186	4.9
Respiratory	222	5.9
<i>Overall, any item missing</i>	<i>1,859</i>	<i>49.5</i>

Table 11 Twenty-five most common missing data patterns

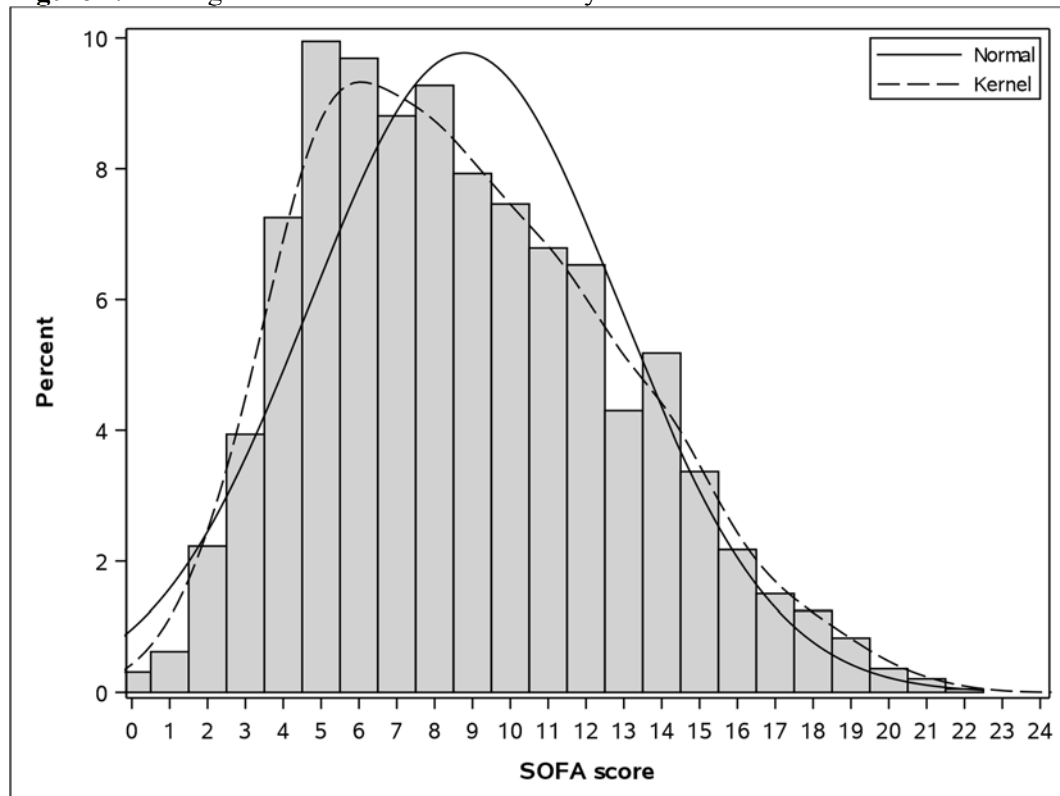
<i>SOFA Score Component</i>								
CNS	Card	Coag	Hep	Ren	Resp	Freq	%	
1	1	1	0	1	1	718	38.6	
0	1	1	1	1	1	495	26.6	
0	1	1	0	1	1	133	7.2	
1	1	1	1	1	0	93	5.0	
1	1	1	0	1	0	41	2.2	
1	1	0	1	1	1	38	2.0	
1	0	1	1	1	1	34	1.8	
1	1	0	0	0	1	33	1.8	
1	1	0	1	0	1	31	1.7	
0	1	1	1	1	0	30	1.6	
1	1	1	0	0	1	17	0.5	
1	1	0	0	1	1	16	0.4	
1	1	1	1	0	1	15	0.4	
0	1	1	0	1	0	15	0.4	
1	0	0	1	0	1	12	0.3	
0	1	0	1	0	1	12	0.3	
1	1	0	0	0	0	11	0.3	
0	0	1	1	1	1	10	0.3	
1	0	1	0	1	1	9	0.2	
1	0	0	0	0	1	8	0.2	
0	1	1	1	0	1	8	0.2	
0	1	0	1	1	1	8	0.2	
0	1	0	0	0	1	8	0.2	
0	0	0	1	0	1	7	0.2	
1	0	0	1	1	1	6	0.2	

Note: The 1/0 elements here represent items within a response vector, where 1 equals a response (the SOFA score element is observed) and 0 represents a non-response (the SOFA score element is missing). The component names are shortened as follows: *CNS* (central nervous system), *Card* (cardiovascular system), *Coag* (coagulation system), *Hep* (hepatic system), *Ren* (renal system), and *Resp* (respiratory system).

4.1.3.3 Distribution of SOFA Scores

The histogram shown below in Figure 4.2 shows the distribution of SOFA scores within the fully-observed dataset. An overlay of the normal and kernel densities is also given, which shows the distribution is right-skewed, with a maximum observed SOFA score of 22 out of 24.

Figure 4.2 Histogram of SOFA scores in the fully-observed dataset



Another method of examining the distribution of the SOFA scores is shown in Table 12 below. Here the scores are binned into four categories. The first category is 0 to 3, which represents a range from no organ derangement (a SOFA score of zero) to at most moderate organ derangement in only one organ system measured by the SOFA score (a SOFA score of 3) or minor organ derangements in more than one organ system.

The second category is 4 to 8, which represents fairly severe organ derangement in up to two organ systems. This group represents most of the patients found within our fully-observed dataset, representing 45.0% of all ICU admissions. The third category is 9 to 11, which represents fairly severe organ derangement and up to three organ systems. The fourth, and final, category is 12 to 22 — which represents patients who could have organ derangement in up to five SOFA score-measured organ systems, and are expected to have, on average, poorer outcomes than patients with lower SOFA scores.

Table 12 Distribution of SOFA scores in the fully-observed dataset

<i>SOFA Score</i>	<i>Freq</i>	<i>%</i>
0-3	137	7.1
4-8	868	45.0
9-11	428	22.2
12-22	497	25.7

4.2 Fully-Observed Dataset Outcomes

All three outcomes that were analyzed in this simulation study were first analyzed using the fully-observed dataset, which contained 1,930 records. These outcomes were analyzed using the same covariates across all three outcomes. The main predictor, *SOFA score*, was used in all regression models, as were the categorical covariates *age* and *race*, and the binary covariate *sex*.

The *age* variable was grouped as follows: 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80 or older; the reference group was 18-29. The *race* variable was comprised of White, Black, and Other/Unknown; the reference group was White. The *Other* category was comprised of Asian, Hawaiian, and Indian/Alaskan. The *sex* variable was a dichotomous variable indicating male or female; the reference group was male. The

SOFA score variable, while used as a continuous variable in the simulations and in most research studies as a risk adjuster, is shown below in four groups for better illustration of effect size and magnitude on the three outcomes. The SOFA score is grouped as previously shown in Table 12: 0-3, 4-8, 9-11, and 12-24.

4.2.1 Outcome 1: Death

A logistic regression model was fit to examine the first outcome of this study, death (a binary indicator variable with 1 representing an in-hospital death, and 0 representing the patient survived the admission). The Hosmer-Lemeshow goodness of fit test showed that the model was a good fit, with a p-value of 0.4490, and a c-statistic of 0.712. The odds ratios and 95% confidence intervals for these predictors are given in Table 13, and graphically in Figure 4.3.

As one can see from Figure 4.3, as the SOFA score increases, the odds of in-hospital death increase in a stepwise fashion when compared with a low SOFA score range of 0-3. The relatively wide confidence interval for the SOFA score group of 12-22 is attributable to the fewer number of patients within this dataset that contain this high score. Likewise, as age of the patient in the ICU increases, so too does the odds of death. Females exhibited 39.8% higher odds of death in comparison to males (OR 1.398, 95% CI 1.142-1.711). Finally—regarding race—Blacks had 25.4% lower odds of death in comparison to Whites (OR 0.746, 95% CI 0.605-0.919), whereas the Other/Unknown group's odds of death were not statistically different from patients who were White.

For this outcome using SOFA score as a continuous variable, how it was used in the simulations, the univariate results showed SOFA score to be statistically-significant predictor of death (OR 1.192, 95% CI 1.162-1.222, $p < 0.0001$). In the adjusted model—adjusting also for age, race, and sex—SOFA score was similarly predictive of death (OR

1.205, 95% CI 1.174-1.237, $p < 0.0001$). The odds ratios for the other predictors were very similar to those shown below in Table 13, where SOFA score was categorized.

Table 13 Odds ratios and 95% confidence intervals for predicting in-hospital death in the fully-observed dataset

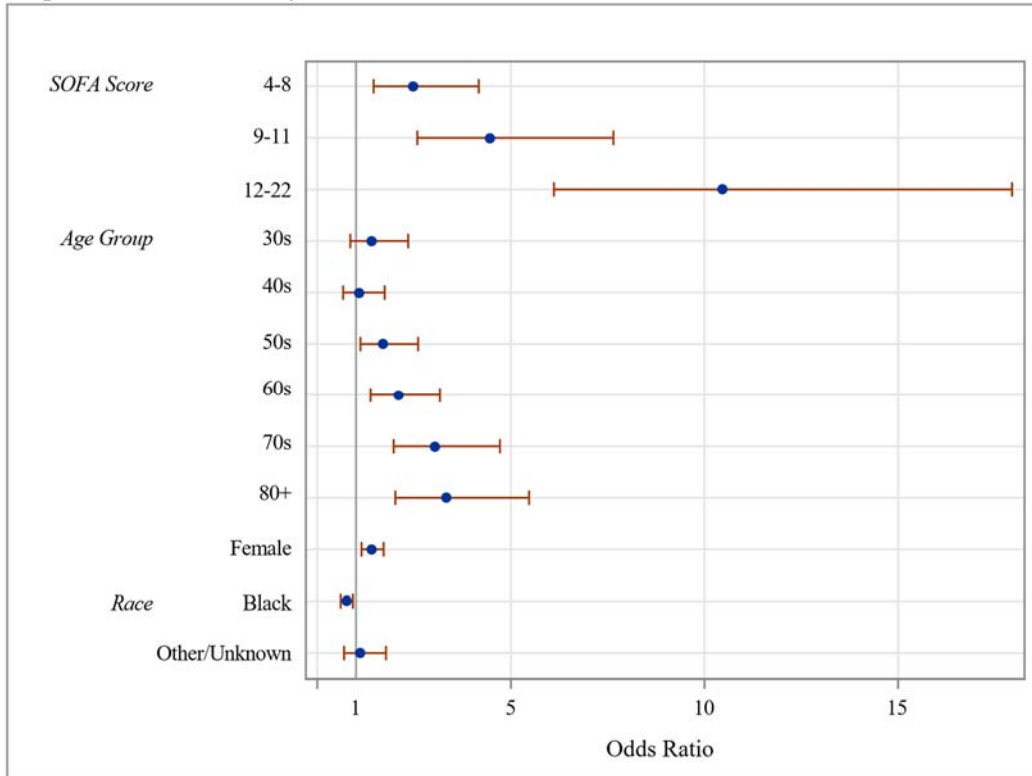
<i>Effect</i>	<i>Odds Ratio</i>	<i>95% Confidence Interval</i>	<i>p-value</i>
SOFA score			
0-3	†		
4-8	2.463	1.455 – 4.170	0.0008
9-11	4.445	2.584 – 7.648	<0.0001
12-24	10.472	6.111 – 17.942	<0.0001
Age group			
18-29	†		
30-39	1.413	0.851 – 2.347	0.1811
40-49	1.077	0.667 – 1.739	0.7608
50-59	1.702	1.114 – 2.602	0.0140
60-69	2.088	1.377 – 3.167	0.0005
70-79	3.050	1.972 – 4.717	<0.0001
80+	3.320	2.015 – 5.469	<0.0001
Female	1.398	1.142 – 1.711	0.0012
Race			
Black	0.746	0.605 – 0.919	0.0060
Other/Unknown	1.109	0.693 – 1.773	0.6672
White	†		

Note. Odds ratio estimates and Wald 95% confidence intervals given. Statistically-significant comparisons at the $\alpha=0.05$ level are given in **bold**.

† indicates the reference group.

Other/Unknown race is comprised of Asian, Hawaiian, Indian/Alaskan, and where this value was missing from the original dataset.

Figure 4.3 Forest plot of the odds ratios and 95% confidence intervals for predicting in-hospital death in the fully-observed dataset



4.2.2 Outcome 2: Total Charges

A gamma-distributed log-linked generalized linear model was fit to examine the second outcome of this study, total charges. This model exhibited good fit, with a deviance of 0.8934 per degree of freedom. In

Table 14 below, the exponentiated least square means estimates and 95% confidence intervals differences of a group compared with a referent group (expressed as a ratio of the reference group) are shown. The exponentiated least squares means estimates of total charges, expressed in 1,000s of dollars, are shown below in Table 15.

Interestingly, these tables show no difference in total charges between SOFA score group 4-8 when compared with the reference group, scores 0-3. However, the total charges for those patients with SOFA scores in the range of 9-11 were 31.64% higher than those in the reference group (ratio 1.316, 95% CI 1.045-1.658), and those with SOFA scores in the range of 12-24 were 36.0% higher than those in the reference group (ratio 1.360, 95% CI 1.148-1.707). This demonstrates the expected behavior that as the SOFA score increases, so do total charges.

For this outcome using SOFA score as a continuous variable, how it was used in the simulations, the univariate results showed SOFA score to be statistically-significant predictor of total charges (β 0.0291, SE 0.0051, $p < 0.0001$). In the adjusted model—adjusting also for age, race, and sex—SOFA score was similarly predictive of total charges (β 0.0255, SE 0.0051, $p < 0.0001$).

Table 14 Differences (expressed as a ratio) between the point estimates and 95% confidence intervals for Total Charges in comparison to reference groups in the fully-observed dataset

<i>Effect</i>	<i>Difference Ratio</i>	<i>95% CI</i>	<i>p-value</i>
SOFA score			
0-3	†		
4-8	1.146	0.975 – 1.346	0.0977
9-11	1.316	1.108 – 1.563	0.0017
12-24	1.360	1.148 – 1.611	0.0004
Age group			
18-29	†		
30-39	0.893	0.741 – 1.075	0.2303
40-49	0.938	0.789 – 1.115	0.4674
50-59	0.820	0.703 – 0.957	0.0119
60-69	0.781	0.670 – 0.910	0.0015
70-79	0.663	0.563 – 0.780	<0.0001
80+	0.553	0.456 – 0.671	<0.0001
Female			
Race			
Black	0.979	0.901 – 1.065	0.6203
Other/Unknown	1.027	0.850 – 1.241	0.7820
White	†		

Note. Differences in total charges, expressed as a ratio from the reference group, are reported along with Wald 95% confidence intervals. Statistically-significant comparisons at the $\alpha=0.05$ level are given in **bold**.

† indicates the reference group.

Other/Unknown race is comprised of Asian, Hawaiian, Indian/Alaskan, and where this value was missing from the original dataset.

Table 15 Least squares means exponentiated point estimates and 95% confidence intervals for Total Charges, expressed in thousands of dollars, in the fully-observed dataset

<i>Effect</i>	<i>Total Charges</i> <i>* \$1,000</i>	<i>95% CI</i>
SOFA score		
0-3	184	169 – 200
4-8	211	191 – 233
9-11	218	198 – 240
12-24	160	137 – 188
Age group		
18-29	242	210 – 279
30-39	216	186 – 250
40-49	227	200 – 258
50-59	199	179 – 220
60-69	189	171 – 209
70-79	160	143 – 180
80+	133	115 – 157
Sex		
Male	197	182 – 213
Female	187	171 – 204
Race		
Black	187	175 – 201
Other/Unknown	197	164 – 236
White	191	179 – 204

Note. Total charges, expressed in 1000s of US dollars (\$), are reported along with Wald 95% confidence intervals. Charges have been rounded to the nearest \$1,000.

Other/Unknown race is comprised of Asian, Hawaiian, Indian/Alaskan, and where this value was missing from the original dataset.

4.2.3 Outcome 3: ICU Length of Stay

A negative binomial-distributed log-linked generalized linear model was fit to examine the third outcome of this study, ICU length of stay. This model exhibited good fit, with a deviance of 1.0625 per degree of freedom. In Table 16 below, the exponentiated least square means estimates and 95% confidence intervals differences of a group compared with a referent group (expressed as a ratio of the reference group) are shown. The exponentiated least squares means estimates of ICU length of stay, expressed in days, are shown below in Table 17.

As expected, patients with higher SOFA scores did exhibit higher ICU lengths of stay on average, as shown in Table 4.9 and 4.10. Interestingly, there were no differences in ICU length of stay amongst the various race categories nor between male and female. However, patients aged 80 or older had lengths of stay in the ICU that were 18.2% shorter in duration than those patients in the reference group, below the age of 30 (difference ratio 0.818, 95% CI 0.673-0.994, $p=0.0432$). Given that this group's odds of in-hospital death were greater than the reference group, the shorter length of stay in the ICU should not be attributable to better outcomes.

For this outcome using SOFA score as a continuous variable, how it was used in the simulations, the univariate results showed SOFA score to not be a statistically-significant predictor of ICU length of stay (β 0.0097, SE 0.0052, $p=0.0642$). In the adjusted model—adjusting also for age, race, and sex—SOFA score was similarly not predictive of total charges (β 0.0081, SE 0.0053, $p=0.1210$).

Table 16 Differences (expressed as a ratio) between the point estimates and 95% confidence intervals for ICU Length of Stay in comparison to reference groups in the fully-observed dataset

<i>Effect</i>	<i>Difference</i>		
	<i>Ratio</i>	<i>95% CI</i>	<i>p-value</i>
SOFA score			
0-3	†		
4-8	1.192	1.011 – 1.404	0.0361
9-11	1.306	1.097 – 1.555	0.0027
12-24	1.201	1.011 – 1.426	0.0373
Age group			
18-29	†		
30-39	0.891	0.739 – 1.074	0.2258
40-49	1.131	0.951 – 1.344	0.1648
50-59	0.987	0.845 – 1.152	0.8650
60-69	1.014	0.870 – 1.181	0.8609
70-79	0.892	0.757 – 1.050	0.1688
80+	0.818	0.673 – 0.994	0.0432
Female	0.965	0.890 – 1.048	0.3974
Race			
Black	1.065	0.979 – 1.158	0.1418
Other/Unknown	1.063	0.879 – 1.285	0.5302
White	†		

Note. ICU length of stay, expressed in days, is reported along with Wald 95% confidence intervals. Statistically-significant comparisons at the $\alpha=0.05$ level are given in **bold**.

† indicates the reference group.

Other/Unknown race is comprised of Asian, Hawaiian, Indian/Alaskan, and where this value was missing from the original dataset.

Table 17 Least squares means exponentiated point estimates and 95% confidence intervals for ICU Length of Stay, expressed in days, in the fully-observed dataset

<i>Effect</i>	<i>Estimate, in days</i>	<i>95% CI</i>
SOFA score		
0-3	8.2	7.0 – 9.6
4-8	9.8	9.0 – 10.6
9-11	10.7	9.7 – 11.8
12-24	9.9	8.9 – 10.9
Age group		
18-29	10.0	8.7 – 11.6
30-39	8.9	7.7 – 10.4
40-49	11.3	10.0 – 12.9
50-59	9.9	8.9 – 11.0
60-69	10.2	9.2 – 11.2
70-79	8.9	7.9 – 10.1
80+	8.2	7.0 – 9.6
Sex		
Female	9.4	8.6 – 10.3
Male	9.8	9.0 – 10.6
Race		
Black	9.8	9.1 – 10.5
Other/Unknown	9.8	8.1 – 11.8
White	9.2	8.6 – 9.8

Note. ICU length of stay, expressed in days, is reported along with Wald 95% confidence intervals.

† indicates the reference group.

Other/Unknown race is comprised of Asian, Hawaiian, Indian/Alaskan, and where this value was missing from the original dataset.

4.3 Aim 1 – Results

The simulation runs were compared using three summary test statistics, yielding the properties of the performance of each of the missing data methods examined. As previously mentioned in *Section 3.8.6–Assessment of Simulation*, these three statistics are (1) relative bias, (2) efficiency, and (3) coverage probability.

The first statistic, **relative bias**, gives the magnitude and direction of bias that a missing data introduces. This statistic is calculated as $\frac{\hat{\beta}_i - \beta}{\beta}$, where β represents the population parameter of the SOFA score from the fully-observed dataset and $\hat{\beta}_i$ represents the parameter estimate for the SOFA score for each simulation run. Ideally, this number will be 0%—meaning no bias exists in the missing data technique incorporated.

The second statistic, **efficiency**, is a simple ratio of the variance of the parameter estimate of the SOFA score. This statistic is calculated as $\frac{\hat{\sigma}_s^2}{\sigma^2}$, where the variance of the parameter estimate of the SOFA score $\hat{\beta}_i$, denoted as $\hat{\sigma}_i^2$, is compared to the variance of the parameter estimate for the SOFA score in the fully-observed dataset β , denoted as σ^2 .

The third statistic, **coverage probability**, is the proportion of simulation runs for each missing data method where the parameter estimate for the SOFA score in the fully-observed dataset β , is contained within the confidence interval for the estimated coefficient $\hat{\beta}_i$. Since $\alpha=0.05$ has already been set for this study, we desire a $1-\alpha = 95\%$ or greater coverage probability. In general, for this alpha level 95% coverage is considered to be ideal.

A discussion of the three outcomes examined in this simulation study follows, using the three aforementioned statistics.

4.3.1 Outcome 1: Death

As mentioned in Section 4.2.1, in the full dataset, the SOFA score was predictive of death in the adjusted model (OR 1.205, 95% CI 1.174-1.237, $p < 0.0001$). The four methods for handling missing data at the composite level for the outcome of death vary in their performance. The coverage probability statistic for these methods at the various percentages of missingness are given in Figure 4.4. Complete case analysis as well as multiple imputation produce results at all percentages of missingness that exceed 95%. However, median imputation quickly has low coverage at 20% or greater missingness for MAR and MNAR right. Zero imputation, the recommended method by the creators of the SOFA score, exhibits poor coverage regardless of missing data mechanism and percent missingness.

The relative bias statistic for these methods at the various percentages of missingness are given in Figure 4.5 (*MAR* missing data mechanism), Figure 4.6 (*MNAR Left* missing data mechanism), Figure 4.7 (*MNAR Middle* missing data mechanism), and Figure 4.8 (*MNAR Right* missing data mechanism). For the MAR missing data mechanism, both median imputation and zero imputation show increasing amounts of bias of the SOFA parameter estimates in the negative direction. Both complete case analysis and multiple imputation show relatively unbiased estimates of the SOFA parameter estimate, however the variance of these estimates increases as the percent of missing data increases. This pattern of increasing variance with increasing percent of missing data is the same for all missing data mechanisms (c.f. Figures 4.5 through 4.8).

The efficiency statistic for these methods at increasing percentages of missingness are given in Figure 4.9 (*MAR* missing data mechanism), Figure 4.10 (*MNAR Left* missing data mechanism), Figure 4.11 (*MNAR Middle* missing data mechanism), and Figure 4.12 (*MNAR Right* missing data mechanism). Figure 4.9 shows that with the *MAR* missing data mechanism, efficiency rapidly increases—as does the spread of efficiency, showing much larger variance in the SOFA parameter estimates in comparison to the true parameter estimates, and large change in variance across simulation runs (as demonstrated by the spread of these estimates). This pattern is repeated for all missing data mechanisms for the complete case analysis method (c.f. Figures 4.9 through 4.12). Across all of the missing data mechanisms and increasing percentages of missing data, MI shows good efficiency—near 1.0—for most of the simulation scenarios.

Figure 4.4 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the logistic regression model predicting *Death* (Aim 1 – Composite Level)

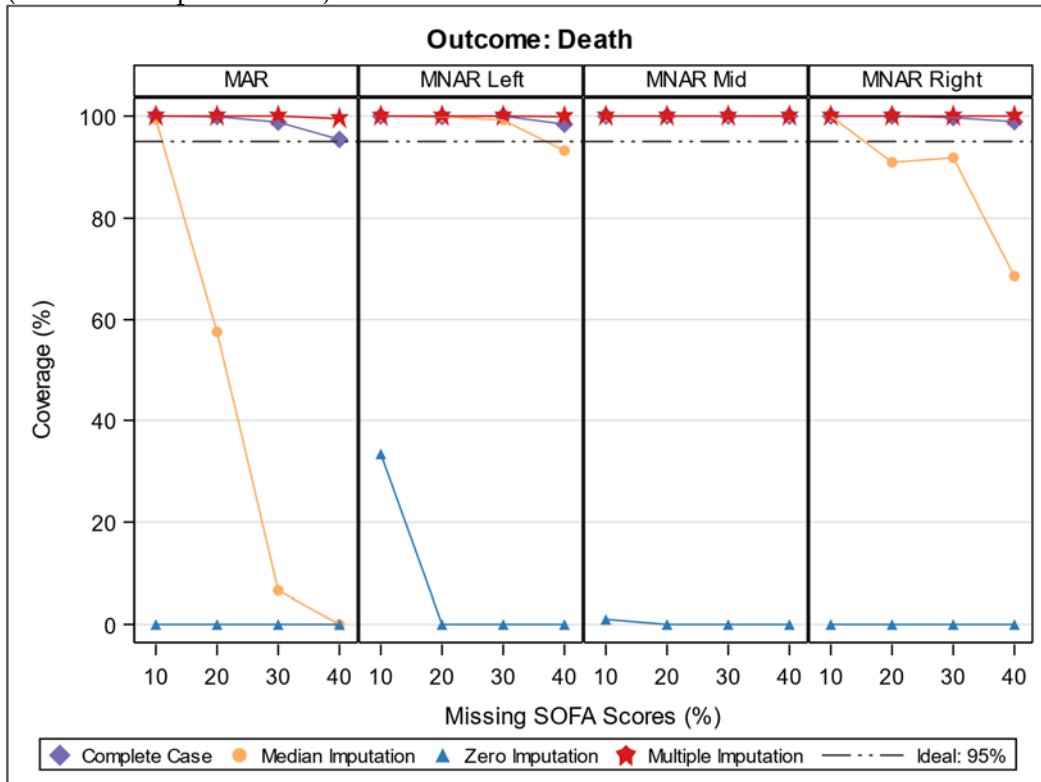


Figure 4.5 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MAR* missing data mechanism (Aim 1 – Composite Level)

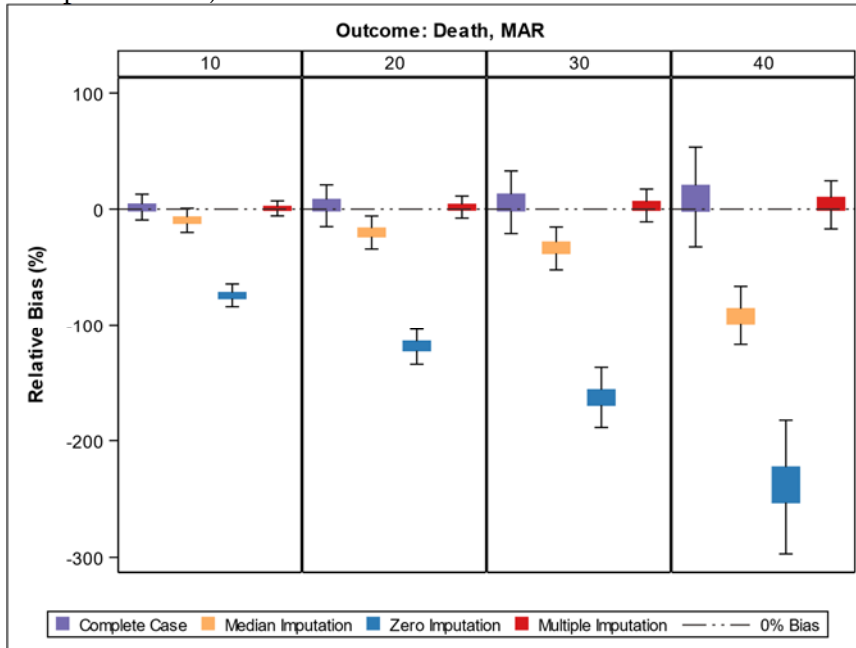


Figure 4.6 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Left* missing data mechanism (Aim 1 – Composite Level)

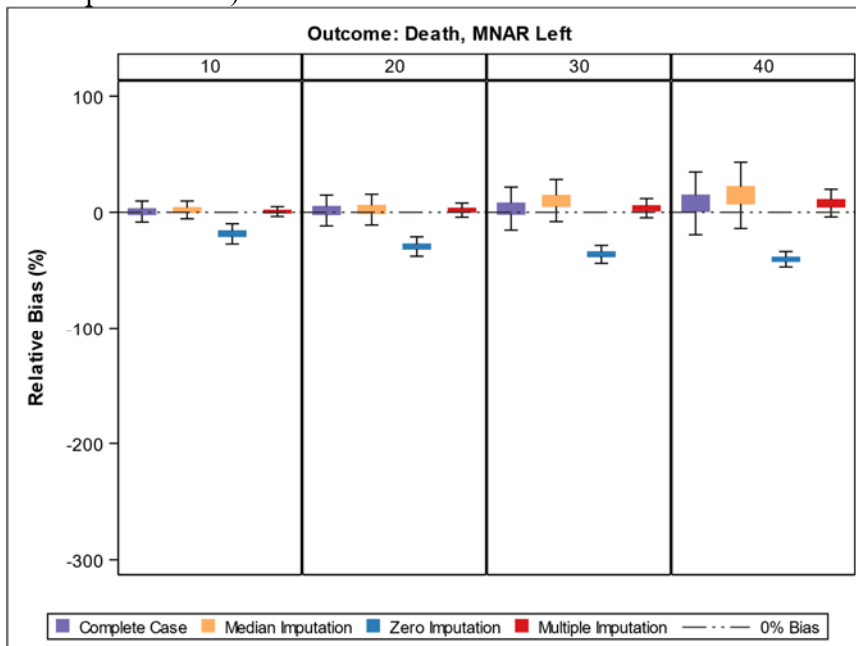


Figure 4.7 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Middle* missing data mechanism (Aim 1 – Composite Level)

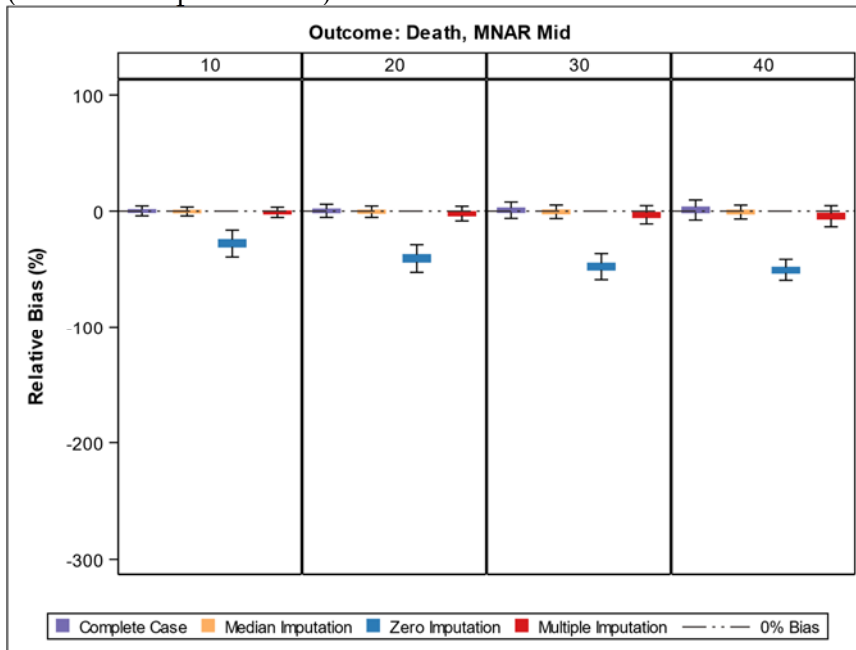


Figure 4.8 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Right* missing data mechanism (Aim 1 – Composite Level)

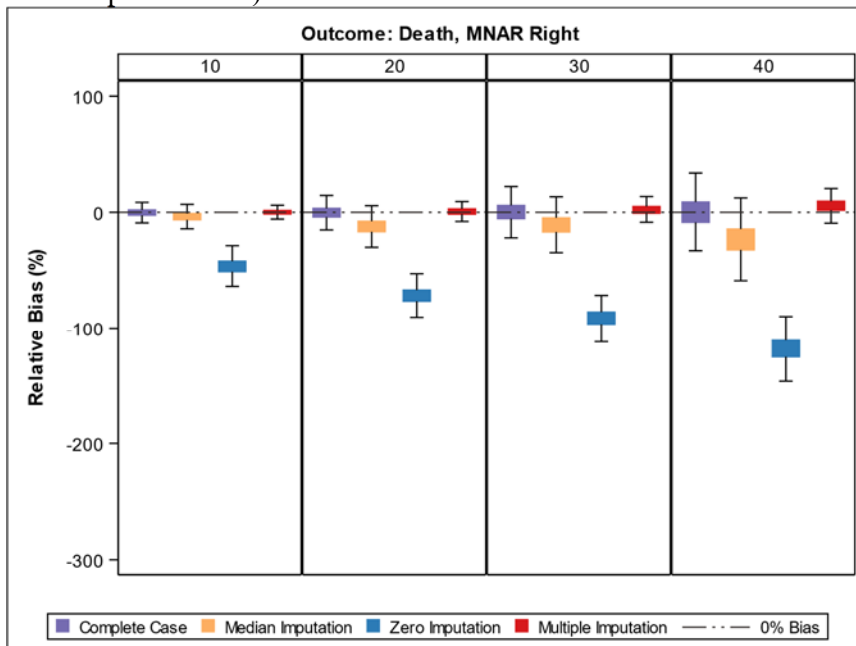


Figure 4.9 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MAR* missing data mechanism (Aim 1 – Composite Level)

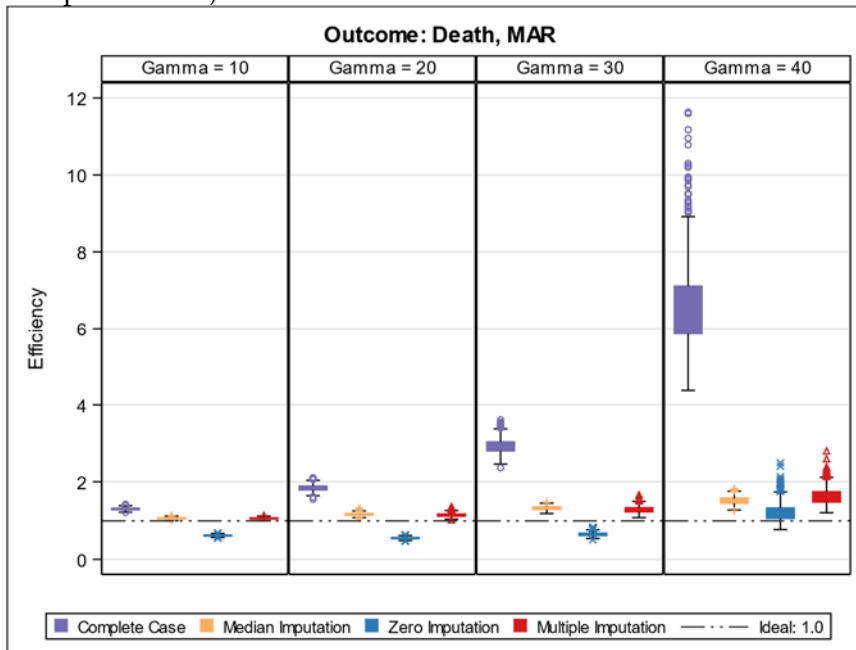


Figure 4.10 Comparison of efficiency estimates for the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Left* missing data mechanism (Aim 1 – Composite Level)

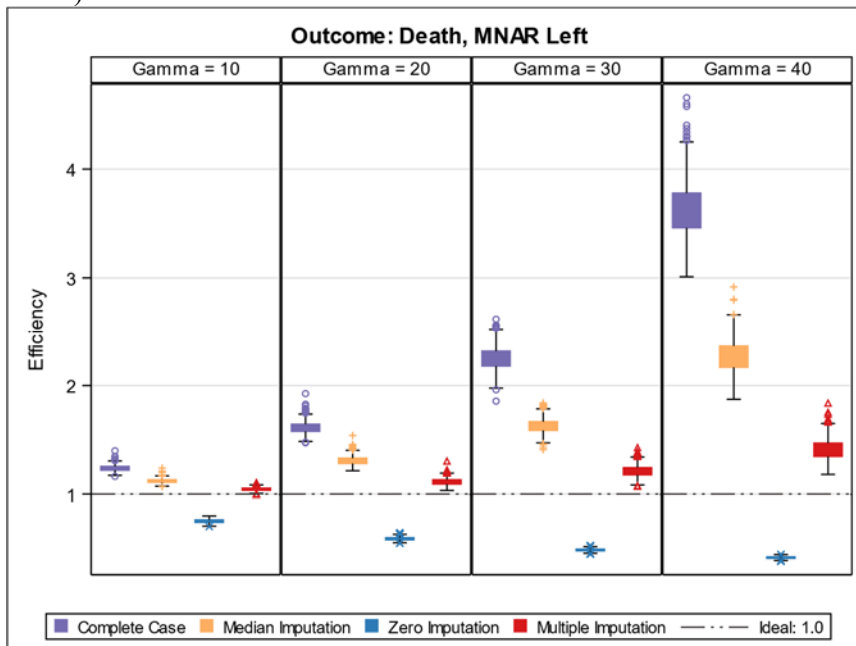


Figure 4.11 Comparison of efficiency estimates for the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Middle* missing data mechanism (Aim 1 – Composite Level)

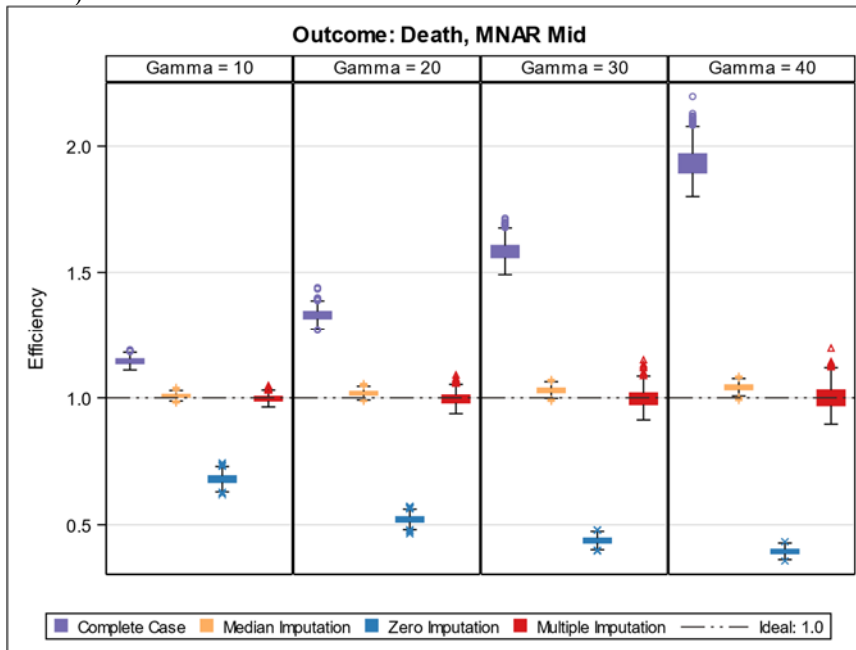
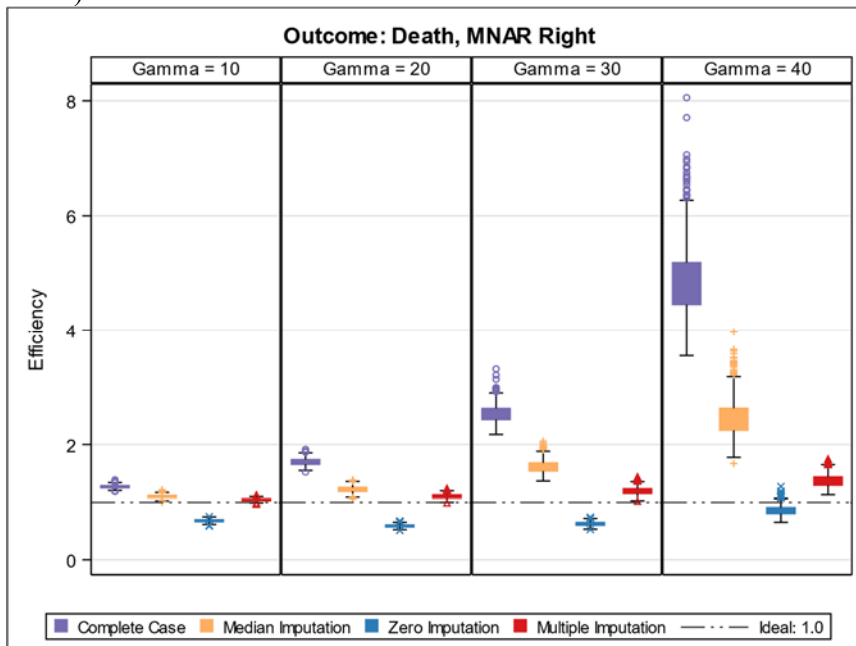


Figure 4.12 Comparison of efficiency estimates for the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Right* missing data mechanism (Aim 1 – Composite Level)



4.3.2 Outcome 2: Total Charges

As mentioned in Section 4.2.2, in the full dataset, the SOFA score was predictive of total charges in the adjusted model (β 0.0255, SE 0.0051, $p < 0.0001$). The four methods for handling missing data at the composite level for the outcome of death vary in their performance. The four methods for handling missing data at the composite level for the outcome of total charges vary in their performance as well. The coverage probability statistic for these methods at the various percentages of missingness are given in Figure 4.13. Multiple imputation produces results at all percentages of missingness that exceed 95%. Complete case analysis and median imputation produced similar results for MNAR middle, but the performance of these methods dropped below 95% for MAR, MNAR left, and MNAR right. Zero imputation, the recommended method by the creators of the SOFA score, exhibited poor coverage regardless of missing data mechanism and percent missingness, with the only exception being the MNAR left mechanism at 10% missingness.

The relative bias statistic for these methods at the various percentages of missingness are given in Figure 4.14 (*MAR* missing data mechanism), Figure 4.15 (*MNAR Left* missing data mechanism), Figure 4.16 (*MNAR Middle* missing data mechanism), and Figure 4.17 (*MNAR Right* missing data mechanism). For the MAR missing data mechanism, both median imputation and zero imputation show increasing amounts of bias of the SOFA parameter estimates in the negative direction. Both complete case analysis and multiple imputation show relatively unbiased estimates of the SOFA parameter estimate, however the variance of these estimates increases as the percent of missing data increases, with complete case having consistently equal or greater variability across simulation runs than multiple imputation. This pattern of

increasing variance with increasing percent of missing data is the same for all missing data mechanisms (c.f. Figures 4.14 through 4.17).

The efficiency statistic for these methods at increasing percentages of missingness are given in Figure 4.18 (*MAR* missing data mechanism), Figure 4.19 (*MNAR Left* missing data mechanism), Figure 4.20 (*MNAR Middle* missing data mechanism), and Figure 4.21 (*MNAR Right* missing data mechanism). Figure 4.18 shows that with the *MAR* missing data mechanism, efficiency rapidly increases—as does the spread of efficiency for all methods besides zero imputation, showing much larger variance in the SOFA parameter estimates in comparison to the true parameter estimates, and large change in variance across simulation runs (as demonstrated by the spread of these estimates). This pattern is repeated for all missing data mechanisms except for the *MNAR middle* mechanism (c.f. Figures 4.18 through 4.21).

In contrast to the other methods, the efficiency rapidly decreases for the zero imputation method, with little variance in these estimates, across all missing data mechanisms (c.f. Figures 4.18 through 4.21). Across all of the missing data mechanisms and increasing percentages of missing data, MI shows good efficiency—near 1.0—for most of the simulation scenarios, albeit with increasing variance as the percentage of missing data increases.

Figure 4.13 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges* (Aim 1 – Composite Level)

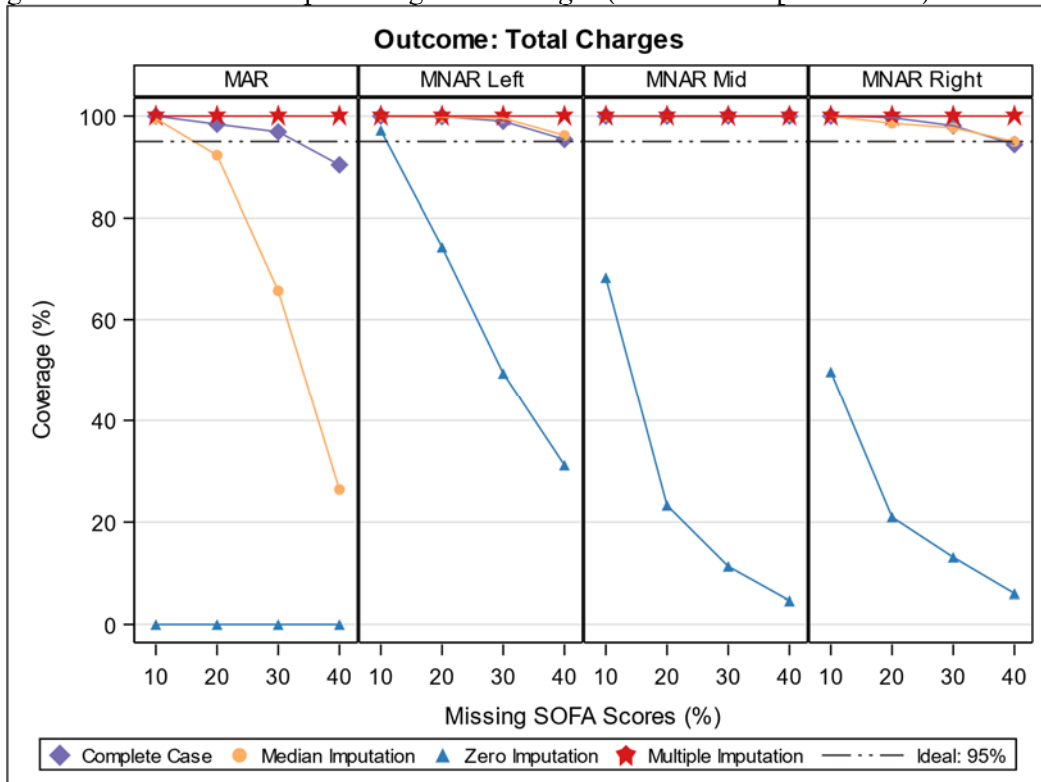


Figure 4.14 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MAR* missing data mechanism (Aim 1 – Composite Level)

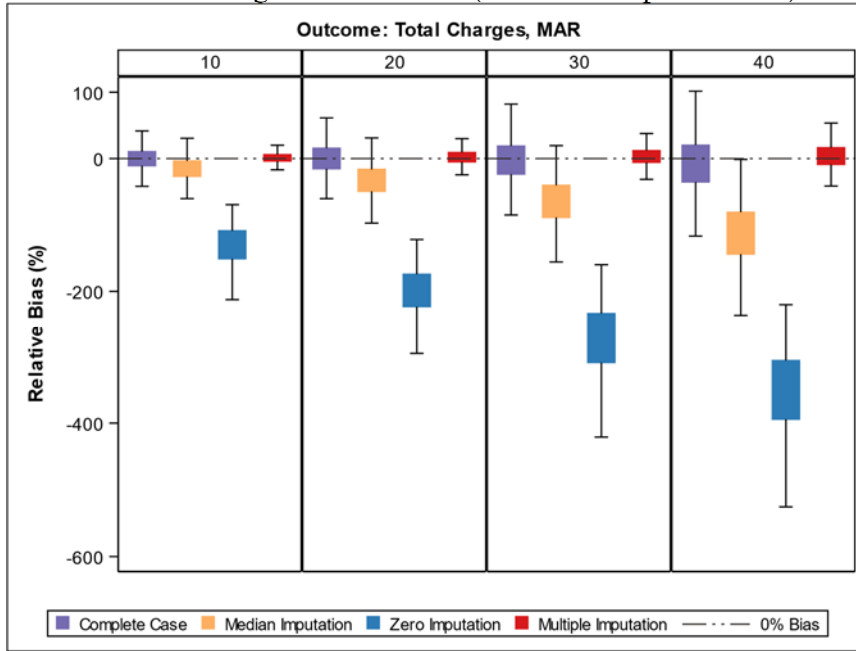


Figure 4.15 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Left* missing data mechanism (Aim 1 – Composite Level)

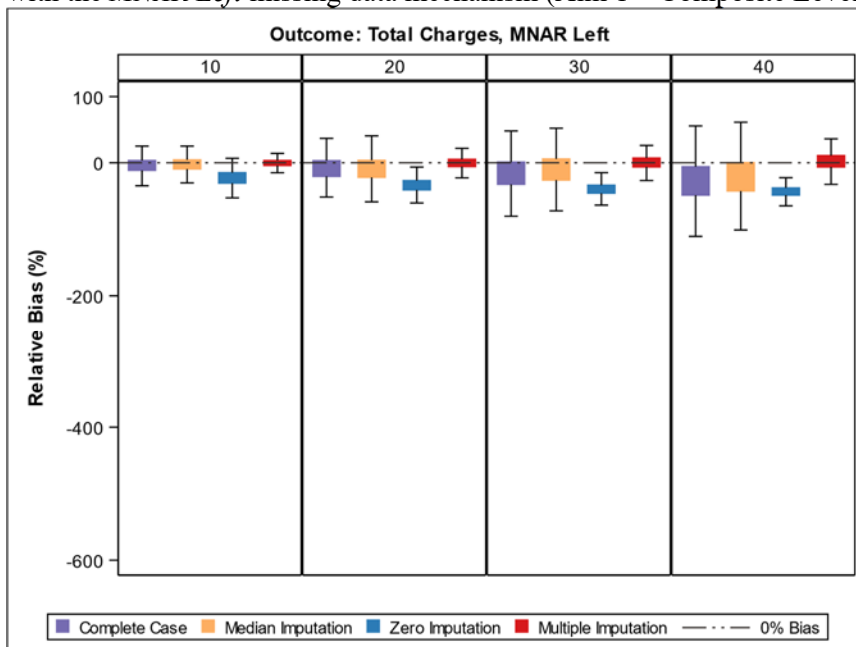


Figure 4.16 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Middle* missing data mechanism (Aim 1 – Composite Level)

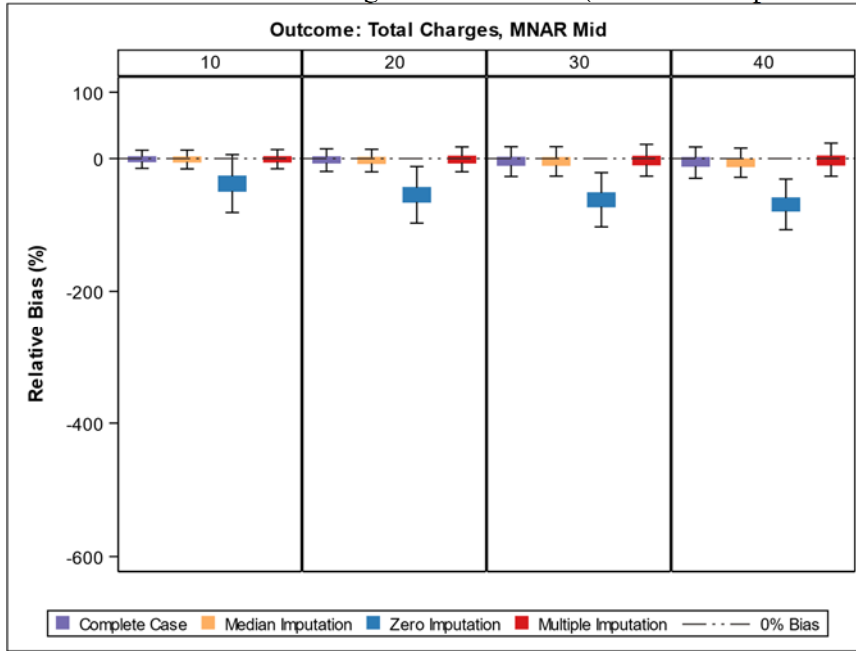


Figure 4.17 Comparison of relative bias of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Right* missing data mechanism (Aim 1 – Composite Level)

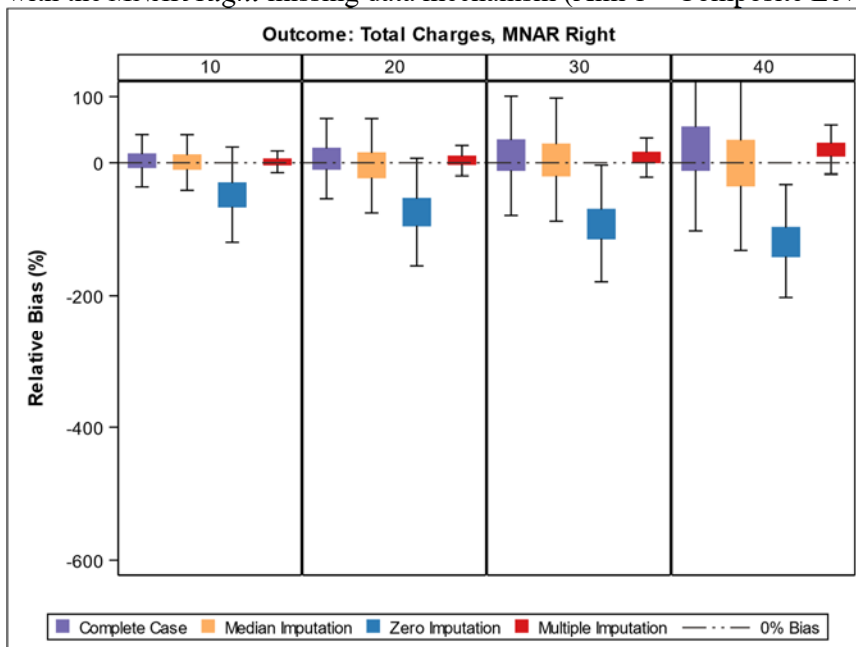


Figure 4.18 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MAR* missing data mechanism (Aim 1 – Composite Level)

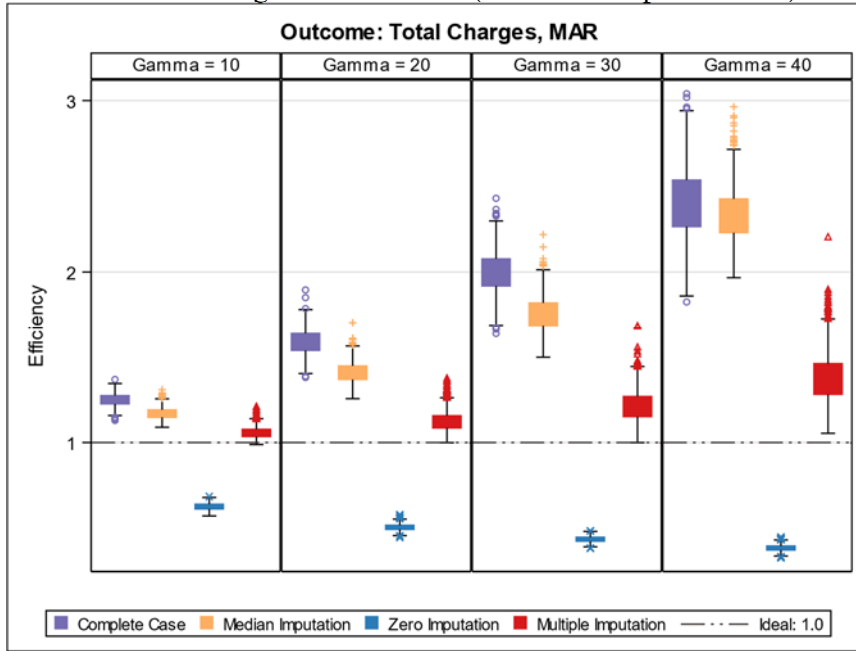


Figure 4.19 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Left* missing data mechanism (Aim 1 – Composite Level)

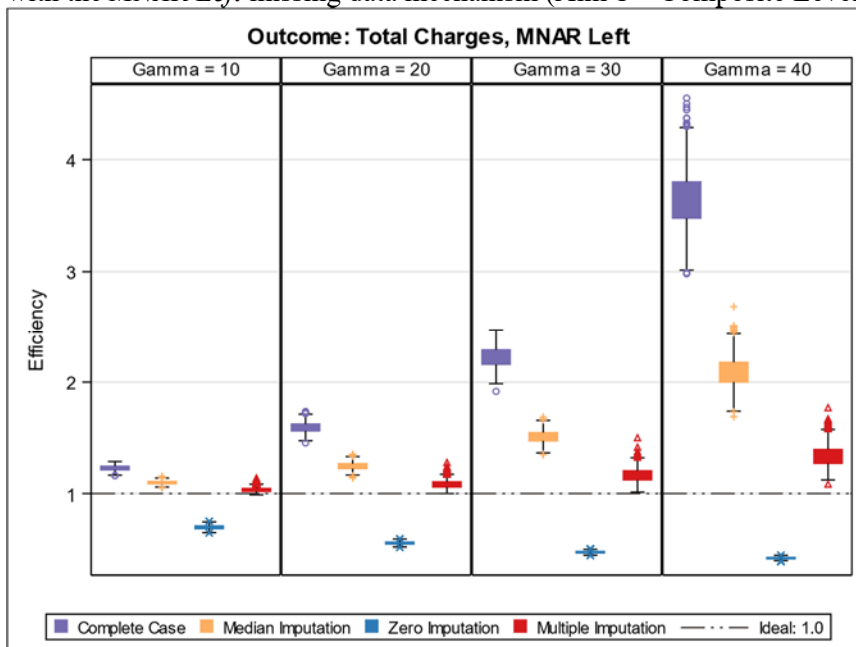


Figure 4.20 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Middle* missing data mechanism (Aim 1 – Composite Level)

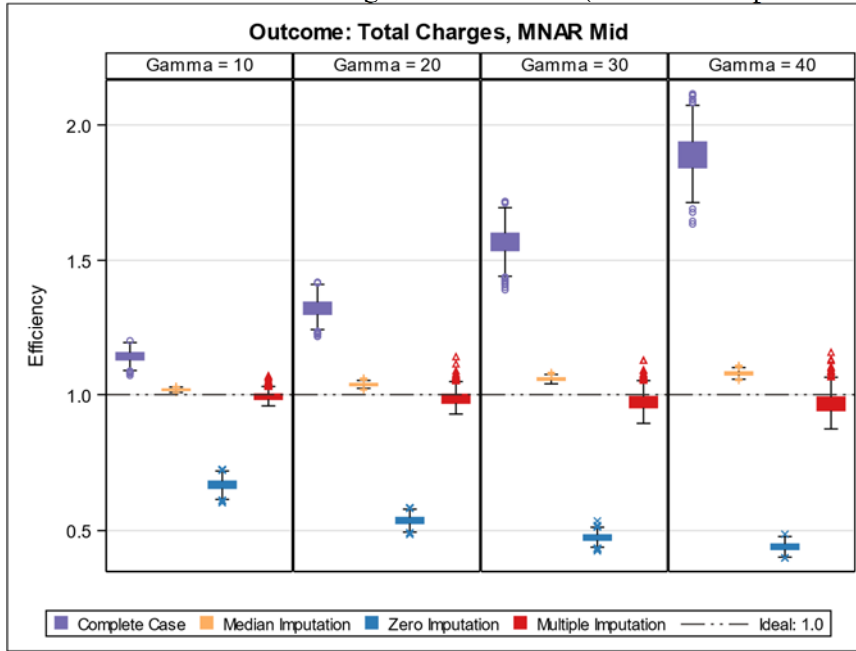
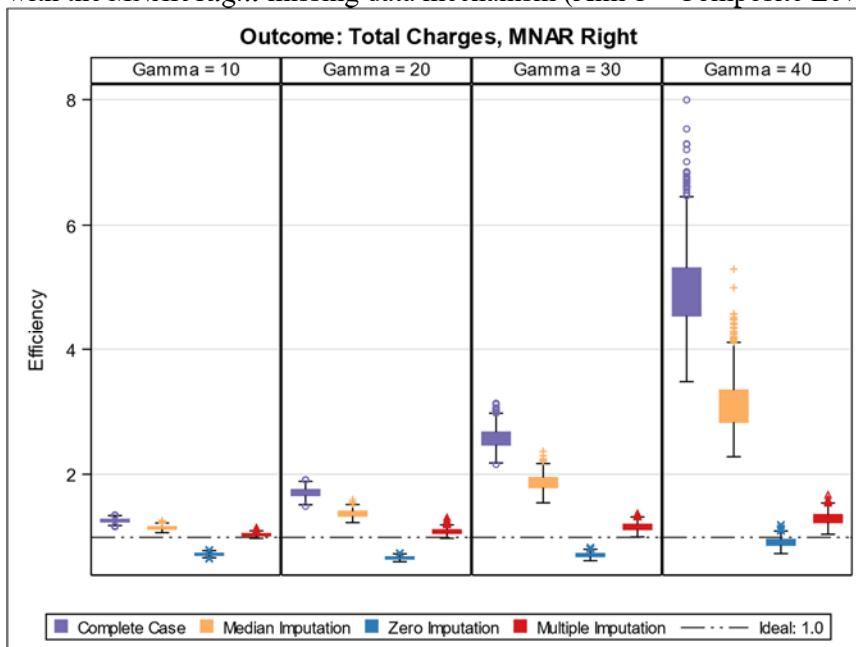


Figure 4.21 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Right* missing data mechanism (Aim 1 – Composite Level)



4.3.3 Outcome 3: ICU Length of Stay

As mentioned in Section 4.2.3, in the full dataset, the SOFA score was not predictive of ICU length of stay in the adjusted model (β 0.0081, SE 0.0053, $p=0.1210$). The four methods for handling missing data at the composite level for the outcome of ICU length of stay vary in their performance as well, as they did for the previous outcomes. The coverage probability statistic for these methods at the various percentages of missingness are given in Figure 4.22. Multiple imputation produced results at all percentages of missingness that exceed 95%. Complete case analysis and median imputation produced similar results for MNAR middle, but the performance of these methods dropped below 95% for MAR, MNAR left, and MNAR right. Zero imputation, the recommended method by the creators of the SOFA score, exhibited poor coverage for all missing data mechanisms and percent missingness, with the only exception being the MNAR left mechanism.

The relative bias statistic for these methods at the various percentages of missingness are given in Figure 4.23 (*MAR* missing data mechanism), Figure 4.24 (*MNAR Left* missing data mechanism), Figure 4.25 (*MNAR Middle* missing data mechanism), and Figure 4.26 (*MNAR Right* missing data mechanism). For the MAR missing data mechanism, both median imputation and zero imputation show increasing amounts of bias of the SOFA parameter estimates in the negative direction. Both complete case analysis and multiple imputation show relatively unbiased estimates of the SOFA parameter estimate, however the variance of these estimates increases as the percent of missing data increases; moreover, the variance of the relative bias on the multiple imputation estimates is smaller than those of the complete case analysis missing data method. This pattern of increasing variance with increasing percent of

missing data is the same for all missing data mechanisms (c.f. Figures 4.23 through 4.26).

The efficiency statistic for these methods at increasing percentages of missingness are given in Figure 4.27 (*MAR* missing data mechanism), Figure 4.28 (*MNAR Left* missing data mechanism), Figure 4.29 (*MNAR Middle* missing data mechanism), and Figure 4.30 (*MNAR Right* missing data mechanism). Figure 4.27 shows that with the *MAR* missing data mechanism, efficiency rapidly increases—as does the spread of efficiency for all methods besides zero imputation and multiple imputation, showing much larger variance in the SOFA parameter estimates in comparison to the true parameter estimates, and large change in variance across simulation runs (as demonstrated by the spread of these estimates). This pattern is repeated for all missing data mechanisms except for the *MNAR* middle mechanism (c.f. Figures 4.27 through 4.30).

In contrast to the other methods, the efficiency rapidly decreases for the zero imputation method, with little variance in these estimates, across all missing data mechanisms except for *MNAR Right* (c.f. Figures 4.27 through 4.30). Across all of the missing data mechanisms and increasing percentages of missing data, MI shows good efficiency—near 1.0—for most of the simulation scenarios, albeit with increasing variance as the percentage of missing data increases.

Figure 4.22 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay* (Aim 1 – Composite Level)

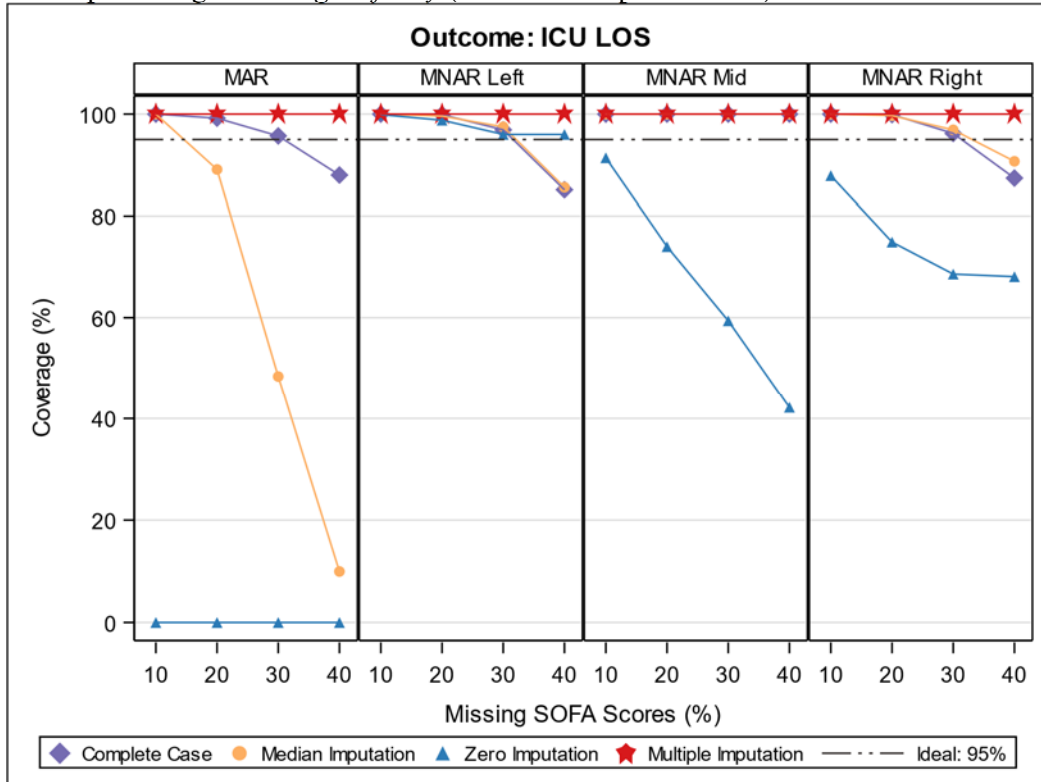


Figure 4.23 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MAR* missing data mechanism (Aim 1 – Composite Level)

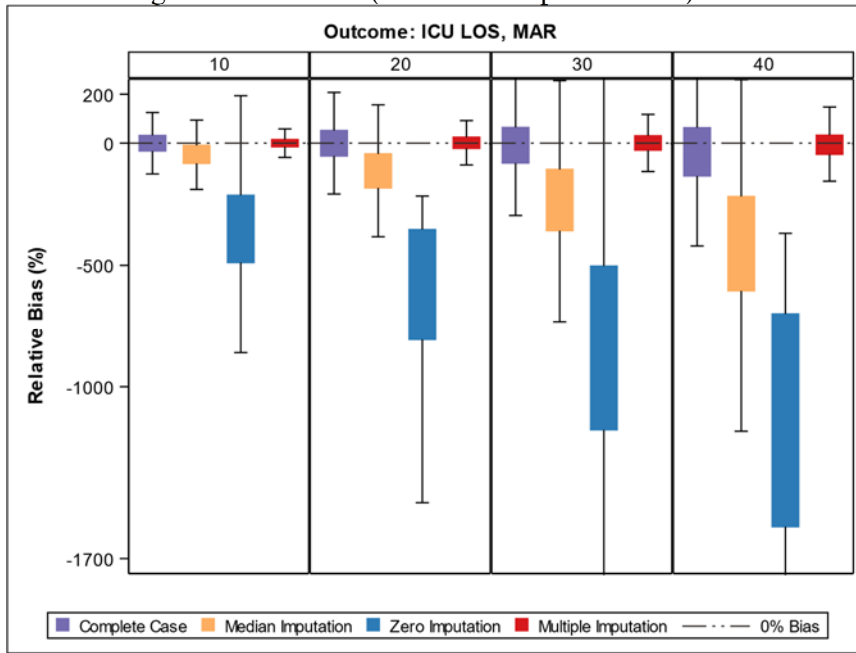


Figure 4.24 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Left* missing data mechanism (Aim 1 – Composite Level)

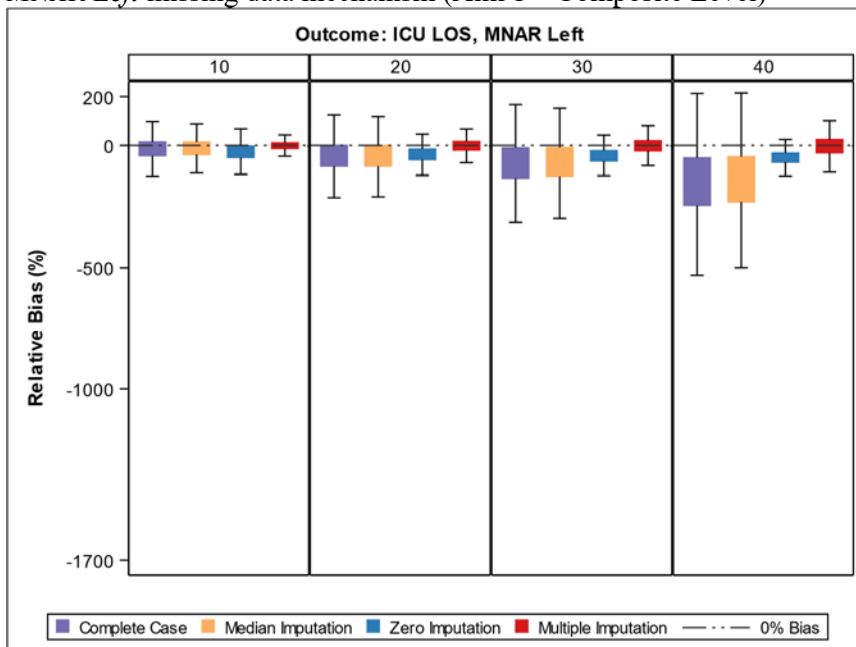


Figure 4.25 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Middle* missing data mechanism (Aim 1 – Composite Level)

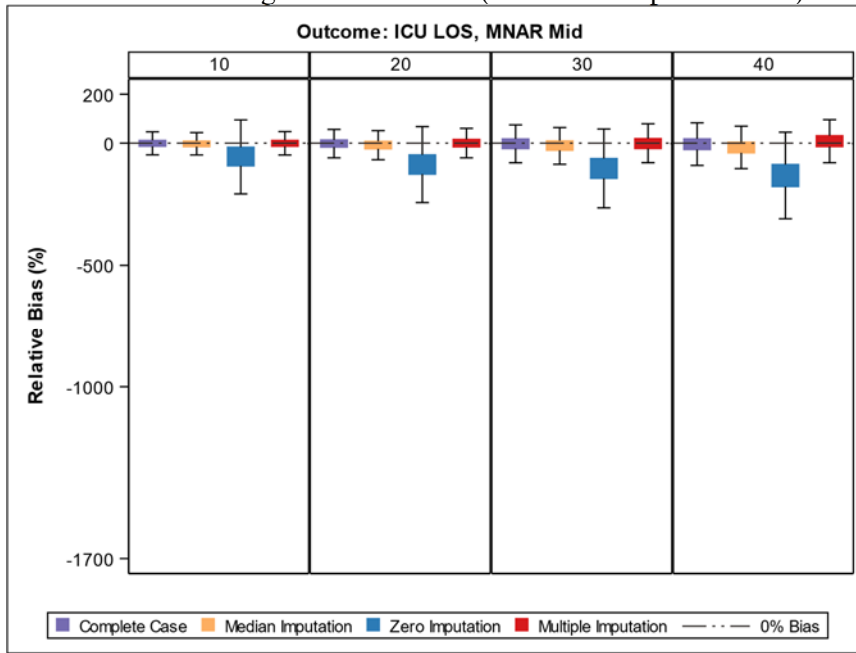


Figure 4.26 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Right* missing data mechanism (Aim 1 – Composite Level)

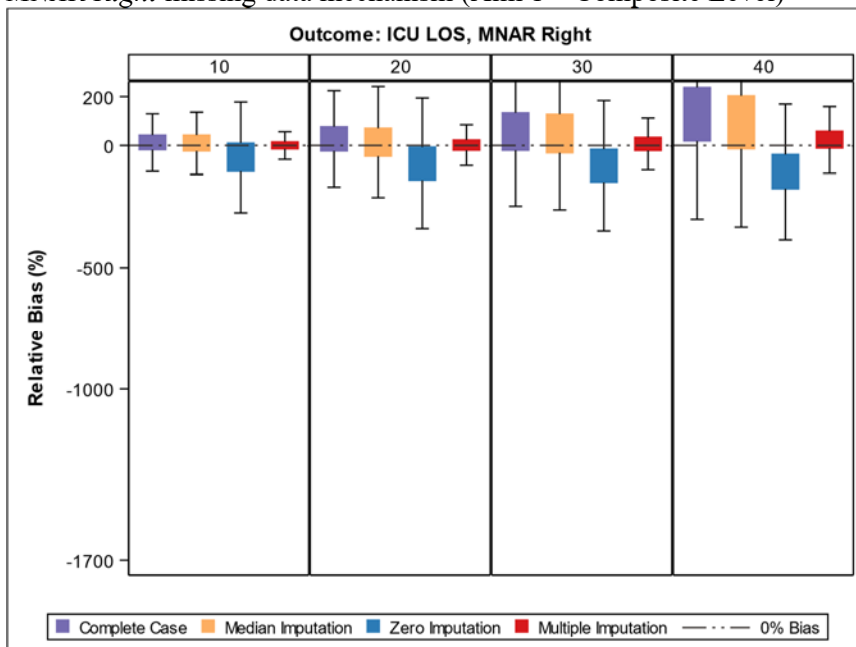


Figure 4.27 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MAR* missing data mechanism (Aim 1 – Composite Level)

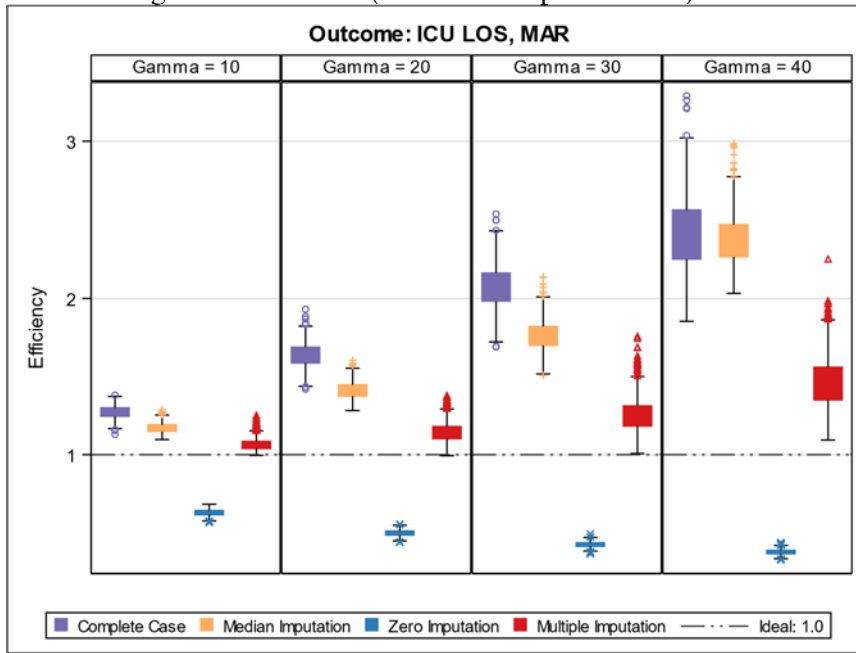


Figure 4.28 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Left* missing data mechanism (Aim 1 – Composite Level)

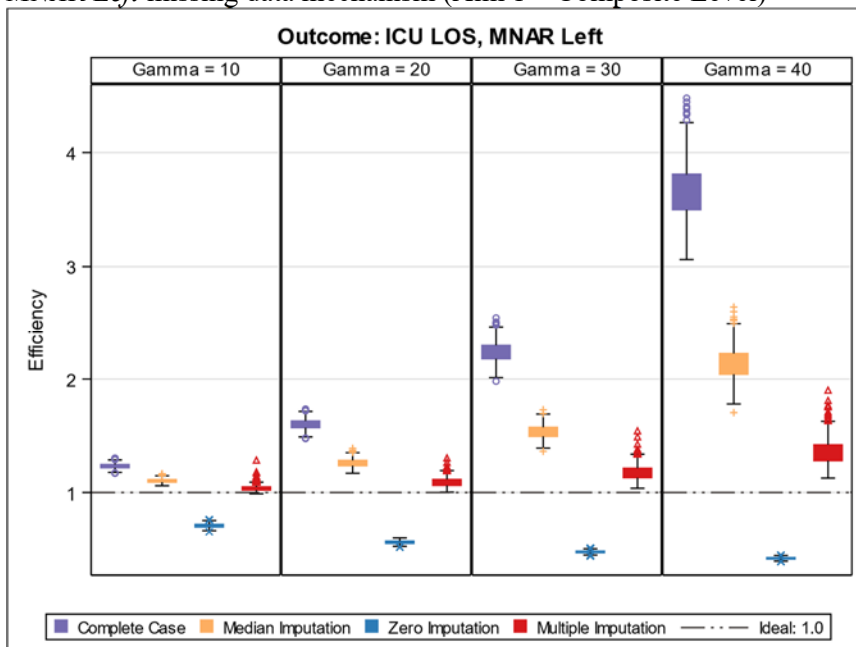


Figure 4.29 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Middle* missing data mechanism (Aim 1 – Composite Level)

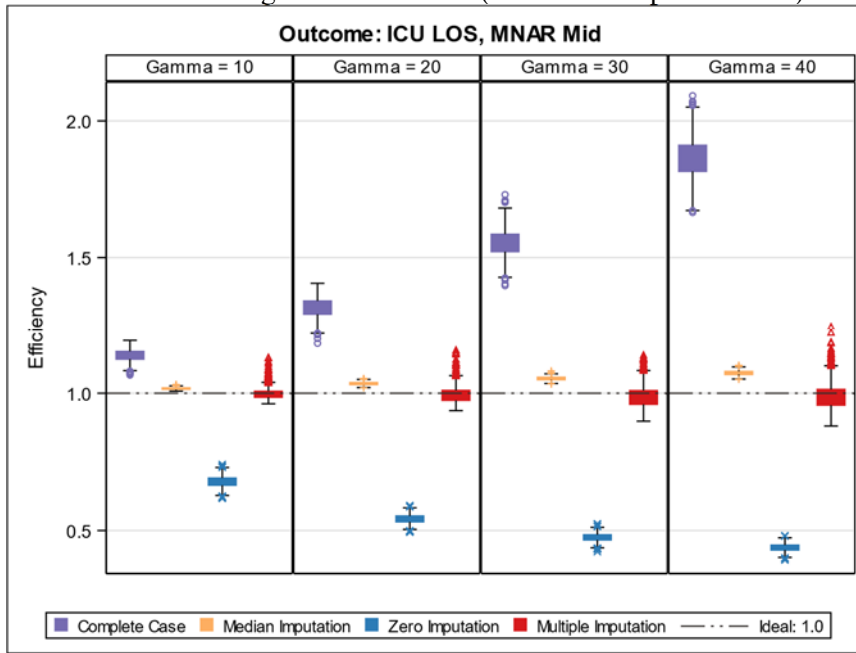
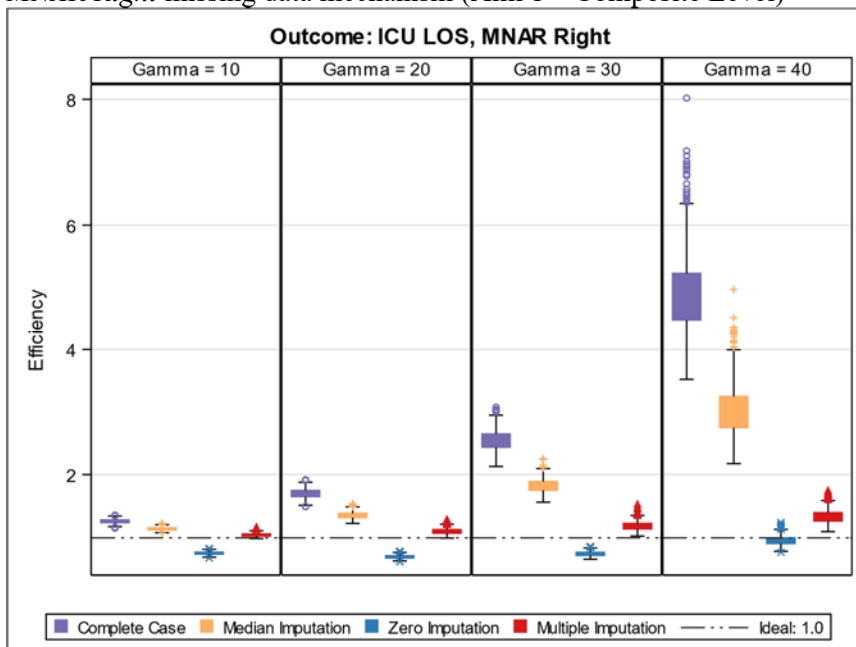


Figure 4.30 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Right* missing data mechanism (Aim 1 – Composite Level)



4.4 Aim 2 – Results

As with Aim 1, these simulation runs were compared using three summary test statistics, yielding the properties of the performance of each of the missing data methods examined. As previously mentioned in *Section 3.8.6–Assessment of Simulation*, these three statistics are (1) relative bias, (2) efficiency, and (3) coverage probability.

4.4.1 Outcome 1: Death

As mentioned in Section 4.2.1, in the full dataset, the SOFA score was predictive of death in the adjusted model (OR 1.205, 95% CI 1.174-1.237, $p < 0.0001$). The four methods for handling missing data at the composite level for the outcome of death vary in their performance. The four methods for handling missing data at the component level for the outcome of death vary in their performance, as they did at the composite level (Aim 1). The coverage probability statistic for these methods at the various percentages of missingness are given in Figure 4.31. Complete case analysis as well as multiple imputation produce results at all percentages of missingness that exceed 95%. Median imputation falls below 95% coverage only for the MAR missing data mechanism at 40% missingness. Zero imputation, the recommended method by the creators of the SOFA score, exhibits poor coverage at 20% or greater missingness with the MAR missing data mechanism, and at 40% missingness with the MNAR left and MNAR right missing data mechanisms.

The relative bias statistic for these methods at the various percentages of missingness are given in Figure 4.32 (*MAR* missing data mechanism), Figure 4.33 (*MNAR Left* missing data mechanism), Figure 4.34 (*MNAR Middle* missing data mechanism), and Figure 4.35 (*MNAR Right* missing data mechanism). For the MAR missing data mechanism, both median imputation and zero imputation show increasing

amounts of bias of the SOFA parameter estimates in the negative direction, with zero imputation exhibiting larger amounts of bias. Both complete case analysis and multiple imputation show relatively unbiased estimates of the SOFA parameter estimate, however the variance of these estimates increases as the percentage of missing data increases. This pattern of increasing variance with increasing percent of missing data is the same for all missing data mechanisms (c.f. Figures 4.32 through 4.35).

The efficiency statistic for these methods at increasing percentages of missingness are given in Figure 4.36 (*MAR* missing data mechanism), Figure 4.37 (*MNAR Left* missing data mechanism), Figure 4.38 (*MNAR Middle* missing data mechanism), and Figure 4.39 (*MNAR Right* missing data mechanism). With all of the missing data mechanisms and at increasing percentages of missingness, the efficiency of the complete case method's parameter estimates rapidly increases, far outpacing the efficiency growth of the other missing data methods (c.f. Figures 4.36 through 4.39). For the other three missing data methods at increasing percentages of missingness, the efficiency stays closer to zero—although the effect is somewhat distorted in these figures, due to scaling.

Figure 4.31 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the logistic regression model predicting *Death* (Aim 2 – Component Level)

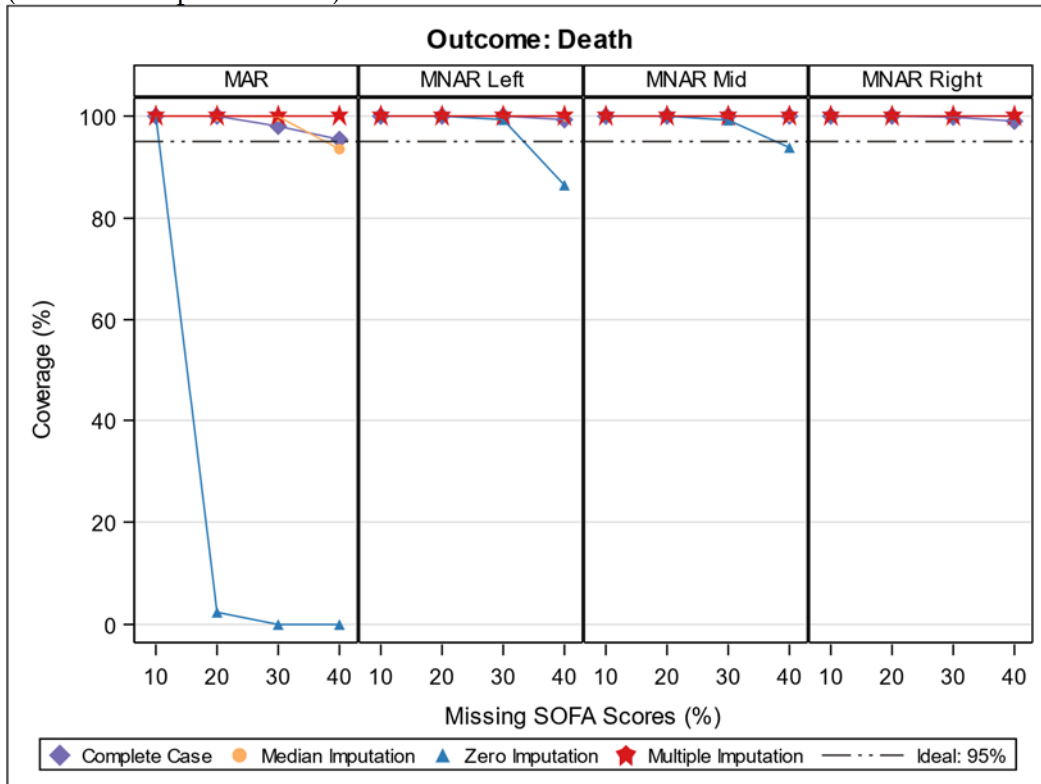


Figure 4.32 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MAR* missing data mechanism (Aim 2 – Component Level)

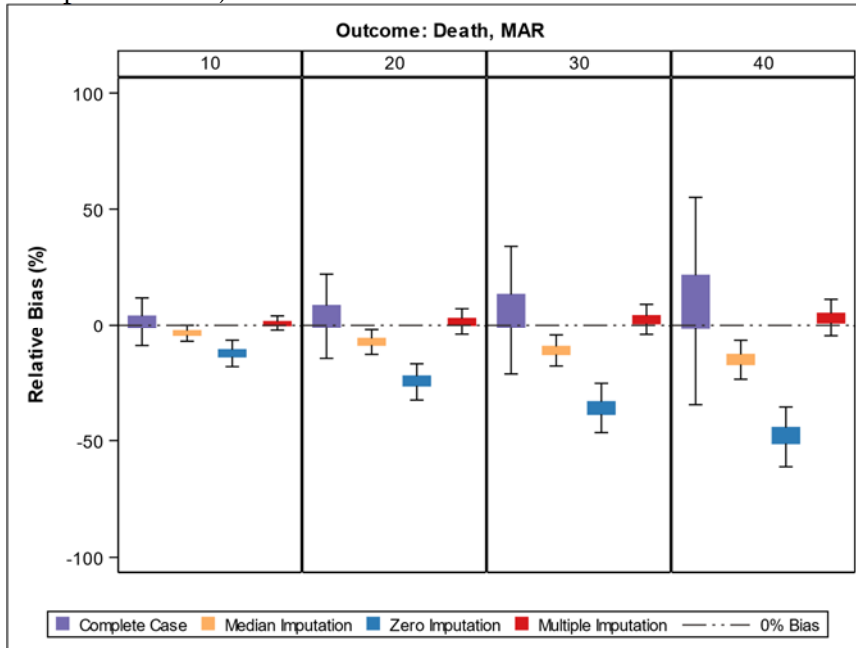


Figure 4.33 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level)

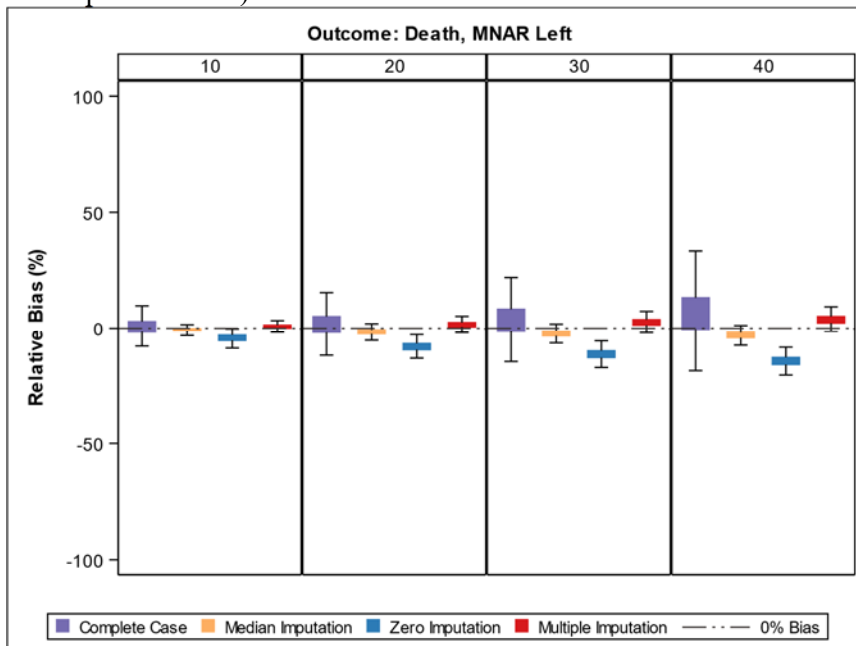


Figure 4.34 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level)

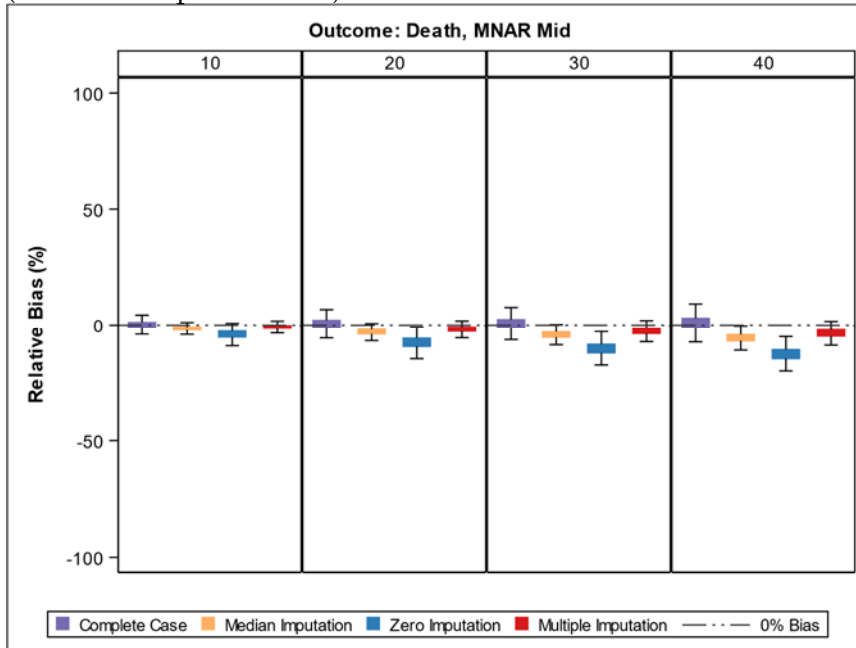


Figure 4.35 Comparison of relative bias of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)

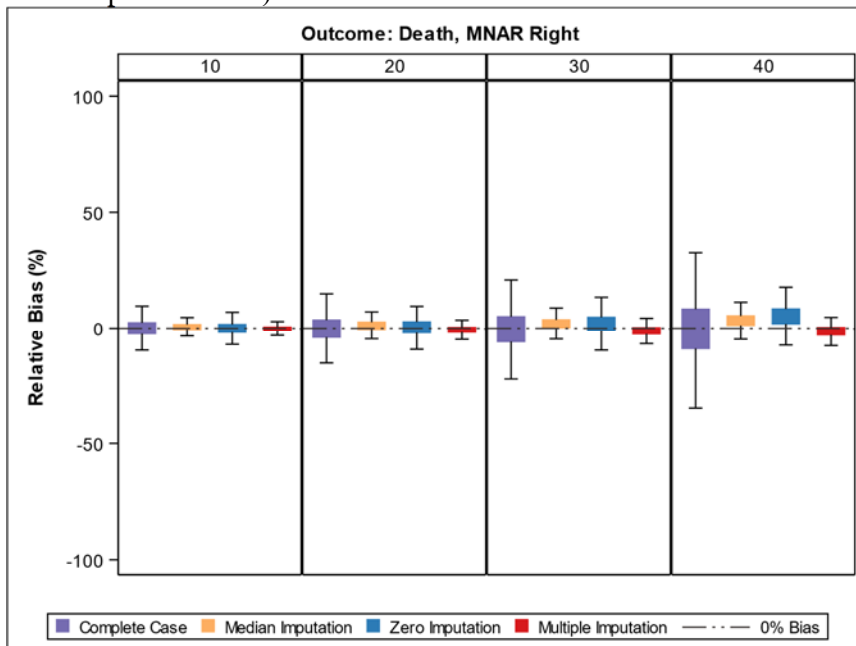


Figure 4.36 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MAR* missing data mechanism (Aim 2 – Component Level)

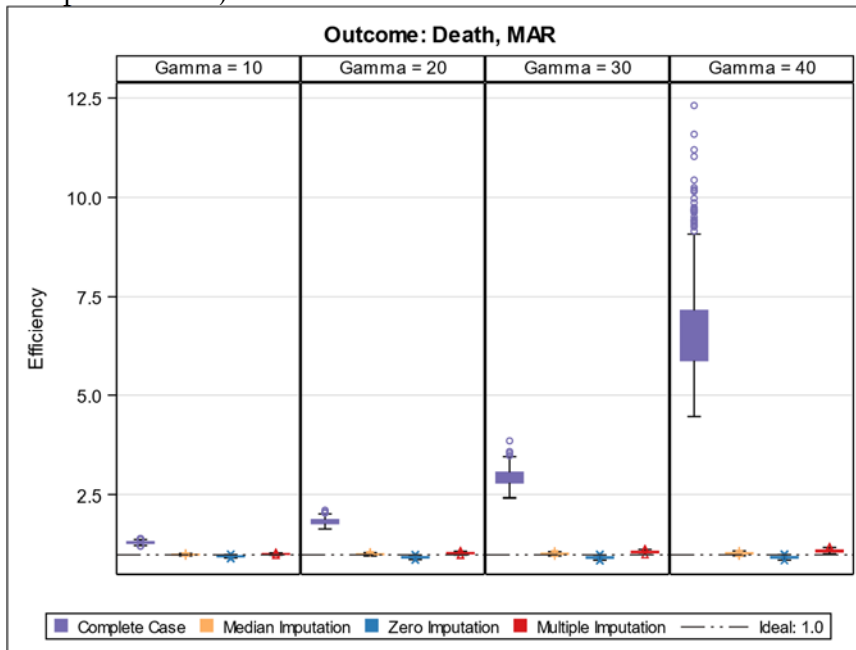


Figure 4.37 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level)

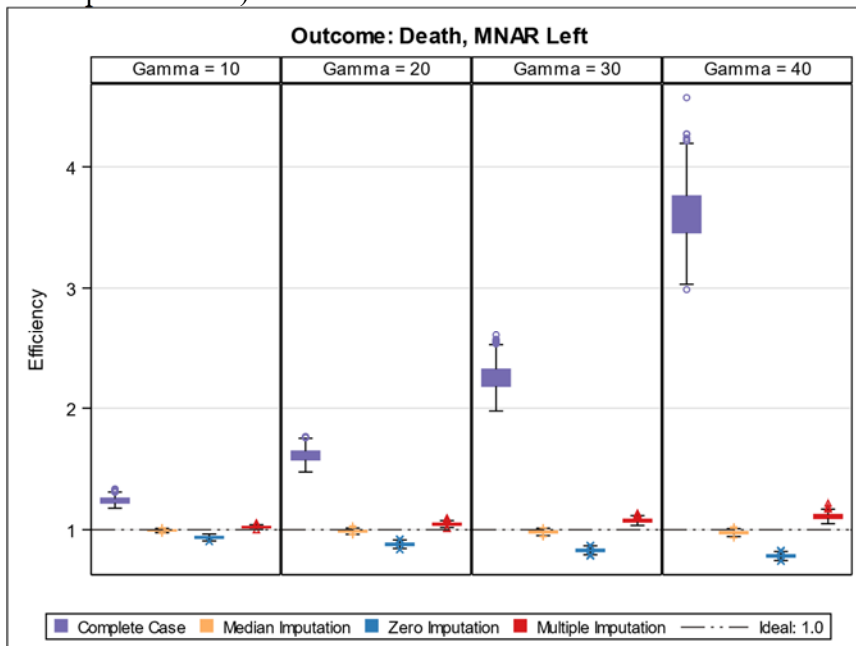


Figure 4.38 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level)

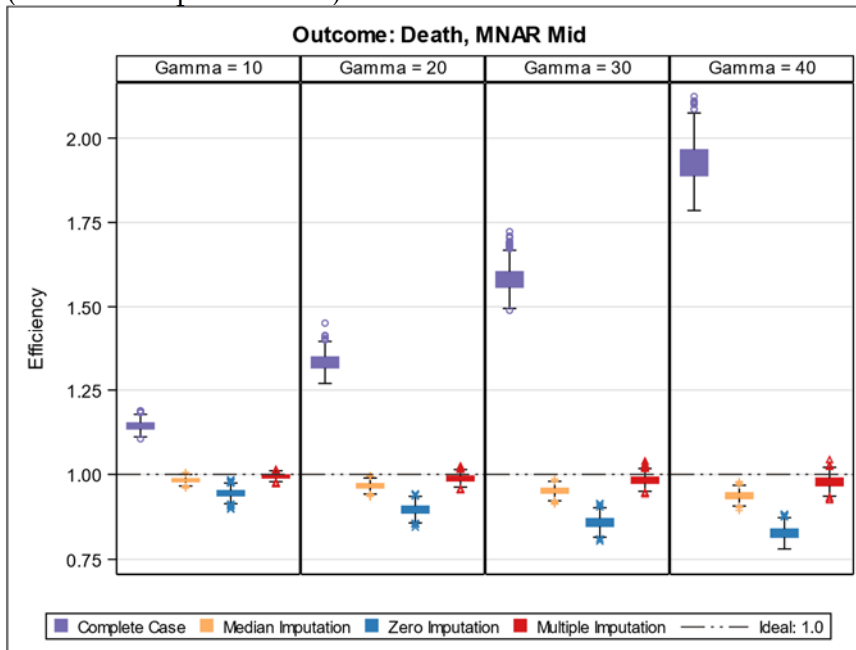
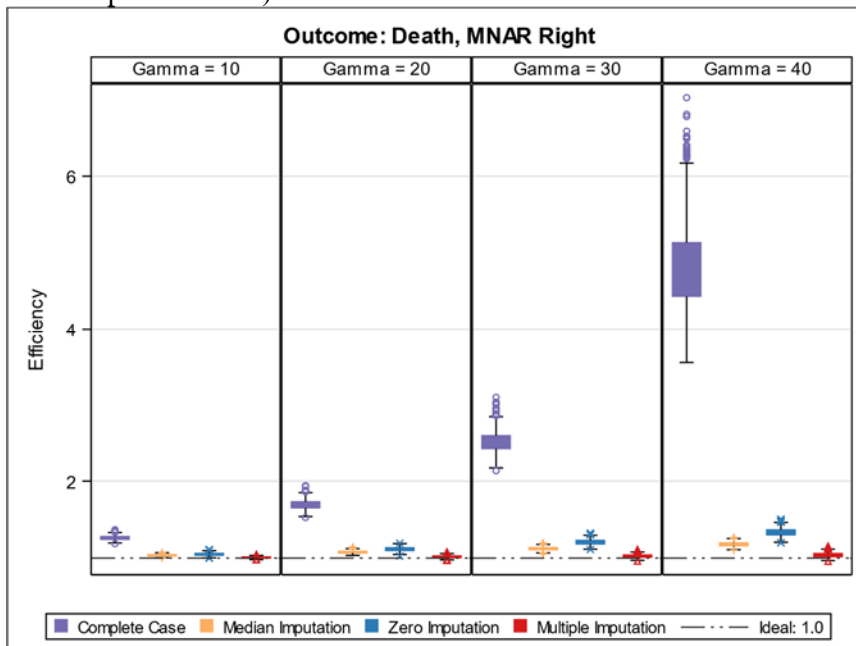


Figure 4.39 Comparison of the efficiency of parameter estimates of the SOFA score among the methods for handling missingness at the composite-level in the logistic regression model predicting *Death*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)



4.4.2 Outcome 2: Total Charges

As mentioned in Section 4.2.2, in the full dataset, the SOFA score was predictive of total charges in the adjusted model (β 0.0255, SE 0.0051, $p < 0.0001$). The four methods for handling missing data at the component level for the outcome of total charges vary in their performance as well, as they did at the composite level (Aim 1). However, the performance for this outcome of handling missing data at the component level (Aim 2) is overall improved over handling missing data at the composite level (Aim 1). The coverage probability statistic for these methods at the various percentages of missingness are given in Figure 4.40. Multiple imputation produced results at all percentages of missingness that exceed 95%. Complete case analysis and median imputation produced similar results for MNAR middle, but the performance of these methods dropped below 95% for MAR, MNAR left, and MNAR right. Zero imputation, the recommended method by the creators of the SOFA score, exhibited poor coverage only for the MAR mechanism at 20% or greater missingness.

The relative bias statistic for these methods at the various percentages of missingness are given in Figure 4.41 (*MAR* missing data mechanism), Figure 4.42 (*MNAR Left* missing data mechanism), Figure 4.43 (*MNAR Middle* missing data mechanism), and Figure 4.44 (*MNAR Right* missing data mechanism). For the MAR missing data mechanism, both median imputation and zero imputation show increasing amounts of bias of the SOFA parameter estimates in the negative direction. Both complete case analysis and multiple imputation show relatively unbiased estimates of the SOFA parameter estimate, however the variance of these estimates increases as the percent of missing data increases, with complete cases being consistently more variable than multiple imputation. However, while the pattern of increasing variance with

increasing percent of missing data is the same for all MNAR missing data mechanisms—MNAR left, MNAR middle, and MNAR right (c.f. Figure 4.42, Figure 4.43, and Figure 4.44, respectively)—most of the methods yielded results that had little or moderate bias.

The efficiency statistic for these methods at increasing percentages of missingness are given in Figure 4.45 (*MAR* missing data mechanism), Figure 4.46 (*MNAR Left* missing data mechanism), Figure 4.47 (*MNAR Middle* missing data mechanism), and Figure 4.48 (*MNAR Right* missing data mechanism). These figures show that with all missing data mechanisms, efficiency rapidly increases—as does the spread of efficiency for the complete case method, showing much larger variance in the SOFA parameter estimates in comparison to the true parameter estimates, and large change in variance across simulation runs (as demonstrated by the spread of these estimates). Similarly, efficiency for the zero imputation method moved away from 1.0 (fully-efficient) as function of increasing percentages of missingness in the negative direction for MNAR left (c.f. Figure 4.46) and MNAR middle (c.f. Figure 4.47), but in the positive direction for MNAR right (c.f. Figure 4.48). Across all of the missing data mechanisms and increasing percentages of missing data, median imputation and MI shows good efficiency—near 1.0—for most of the simulation scenarios, albeit with increasing variance as the percentage of missing data increases.

Figure 4.40 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges* (Aim 2 – Component Level)

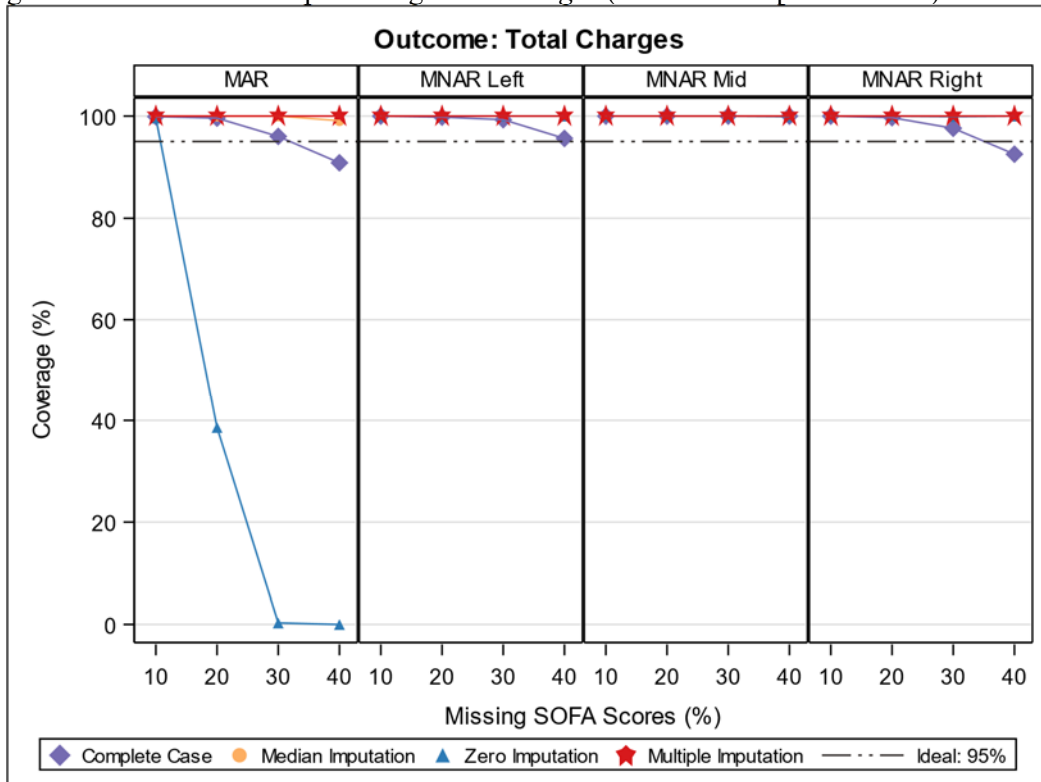


Figure 4.41 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MAR* missing data mechanism (Aim 2 – Component Level)

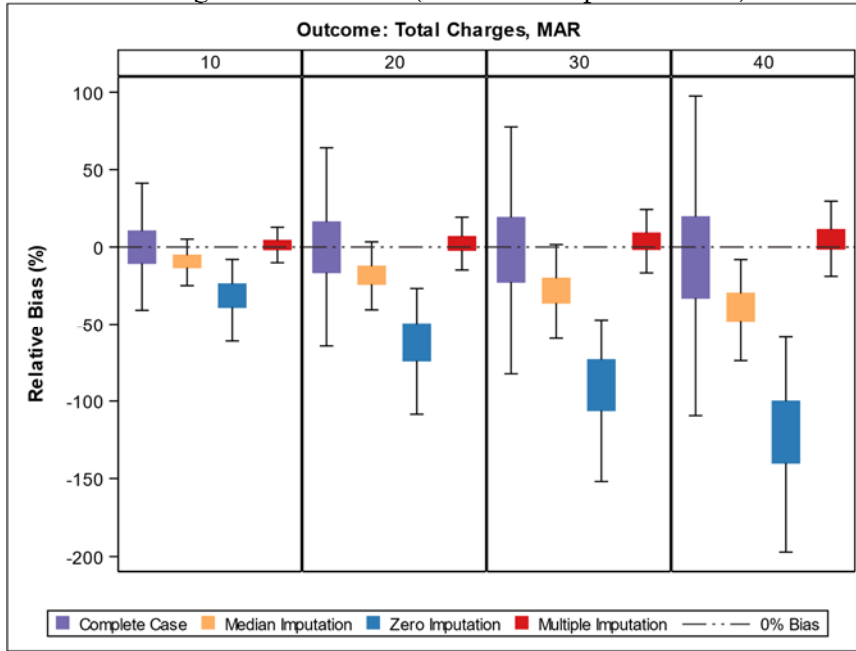


Figure 4.42 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level)

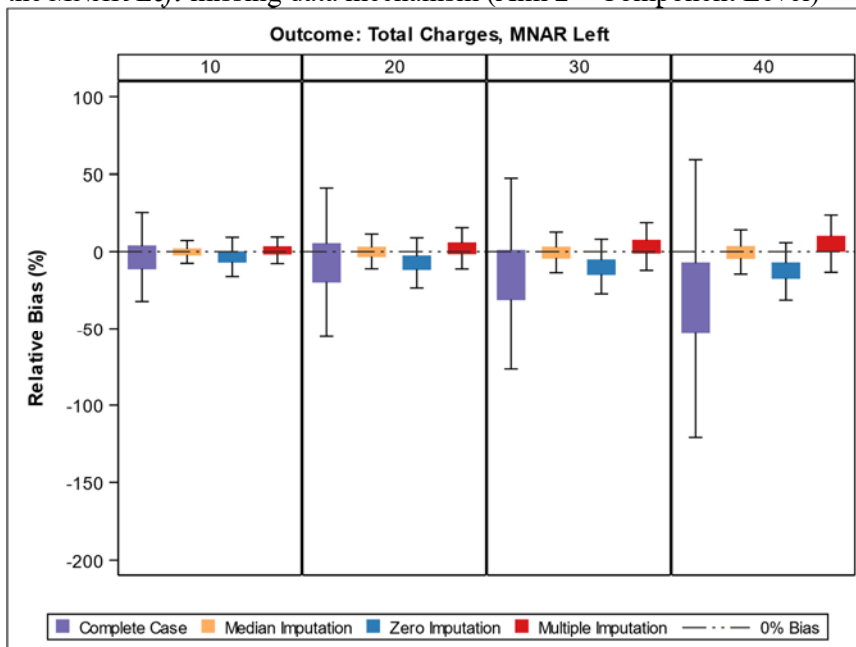


Figure 4.43 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level)

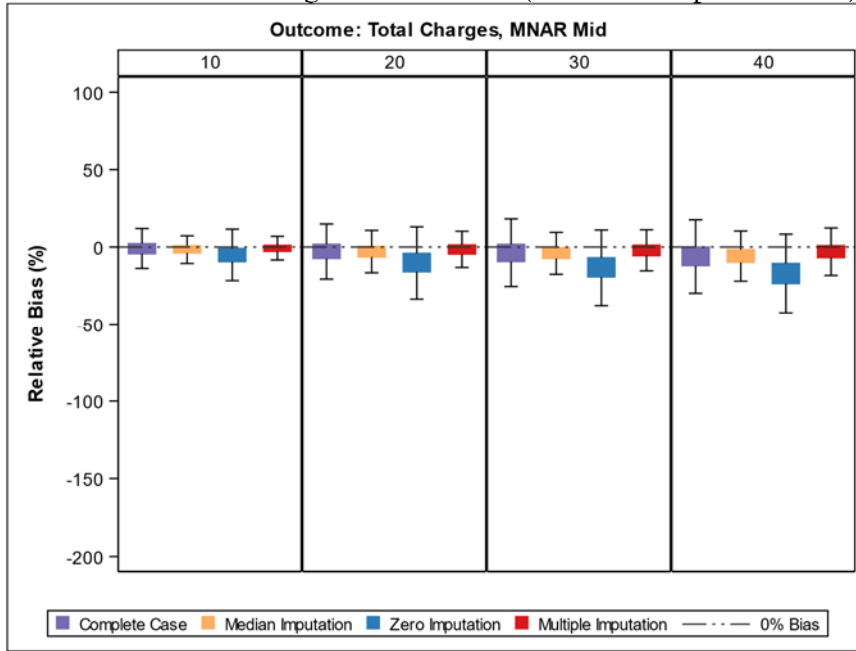


Figure 4.44 Comparison of relative bias of parameter estimates of the SOFA in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)

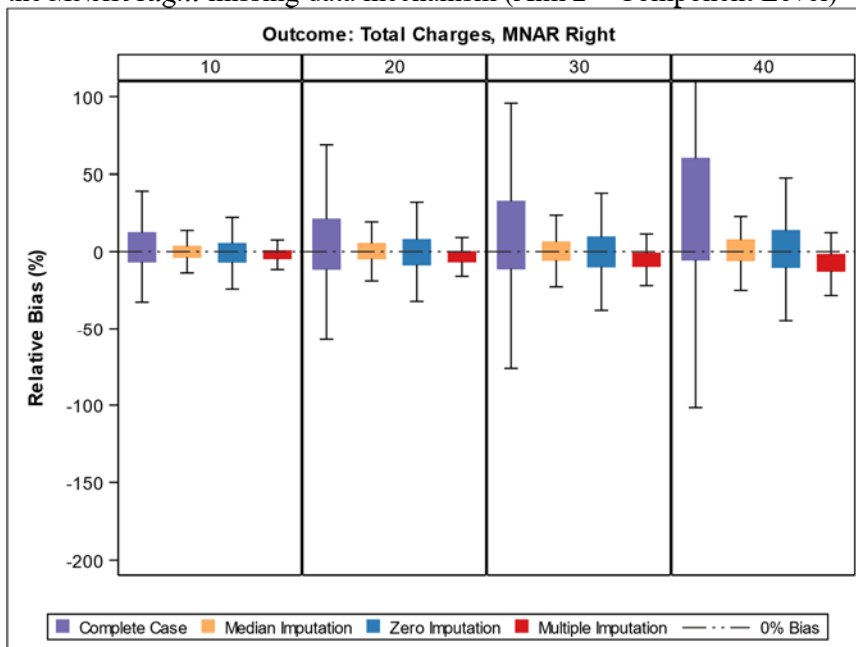


Figure 4.45 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MAR* missing data mechanism (Aim 2 – Component Level)

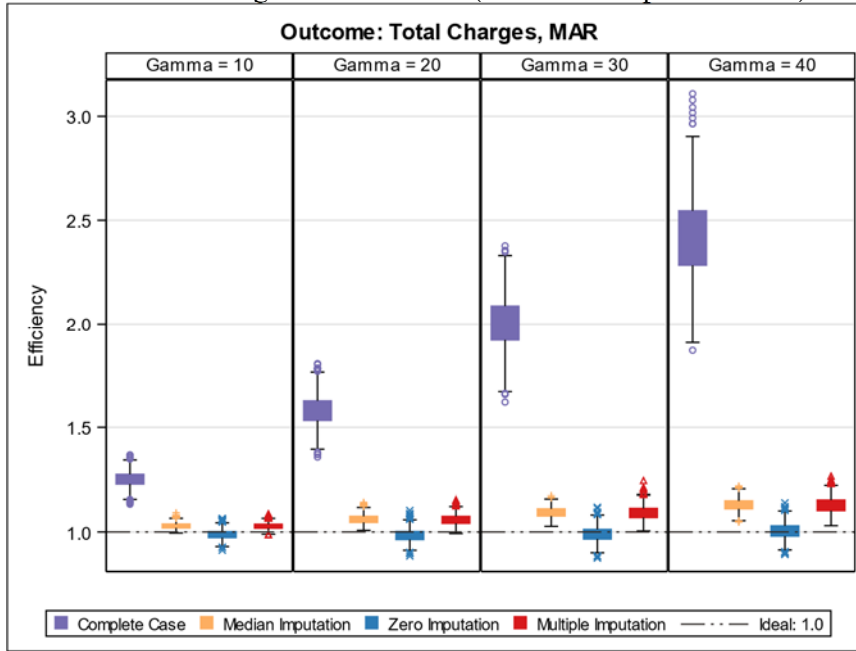


Figure 4.46 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level)

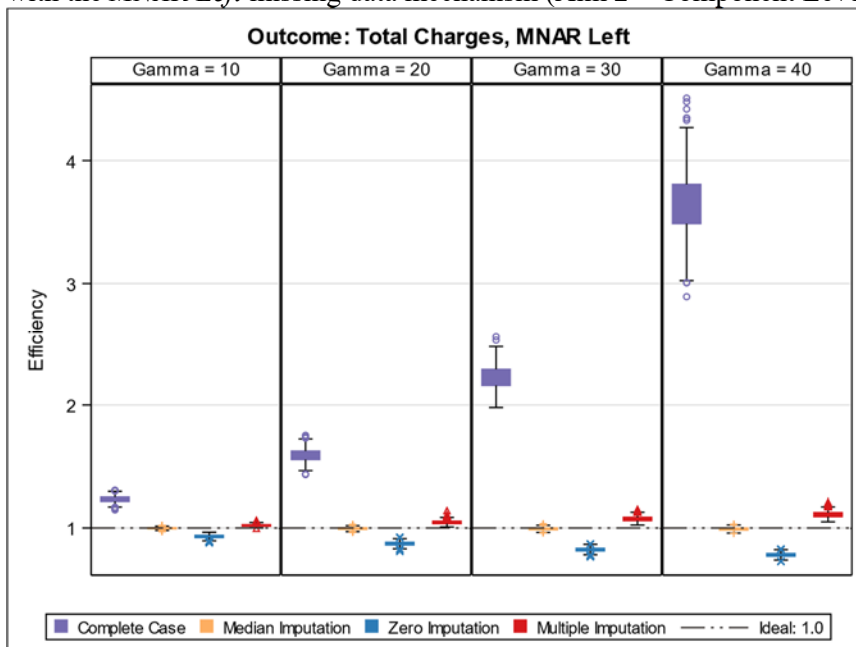


Figure 4.47 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level)

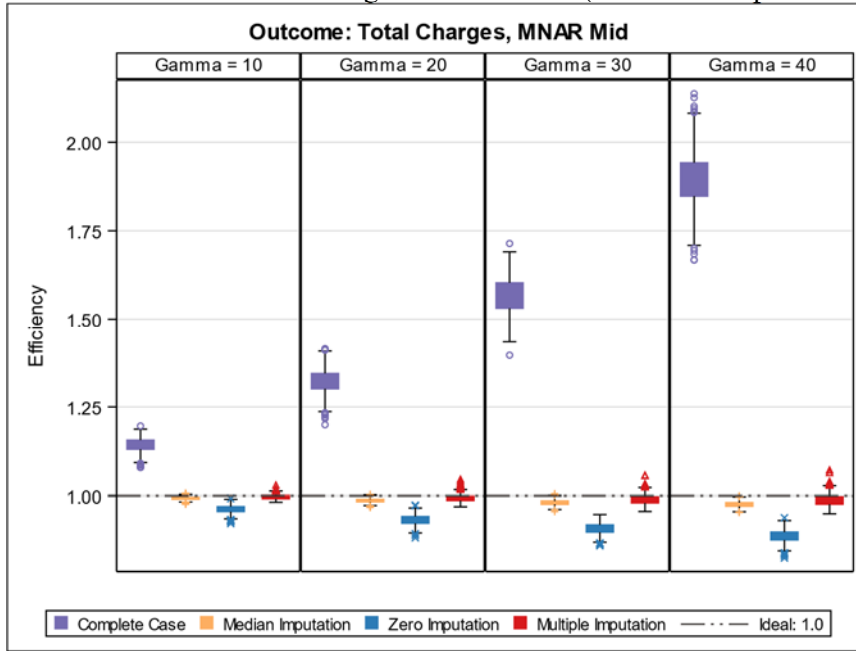
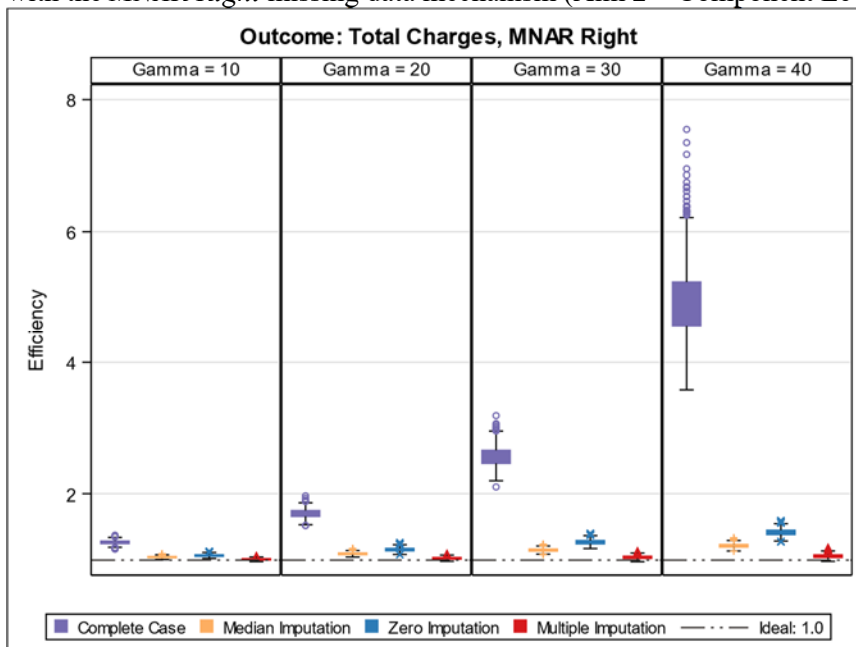


Figure 4.48 Comparison of the efficiency of parameter estimates of the SOFA score in the gamma-transformed log linked generalized linear model predicting *Total Charges*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)



4.4.3 Outcome 3: ICU Length of Stay

As mentioned in Section 4.2.3, in the full dataset, the SOFA score was not predictive of ICU length of stay in the adjusted model (β 0.0081, SE 0.0053, $p=0.1210$). The four methods for handling missing data at the component level for the outcome of ICU length of stay vary in their performance as well, as they did at the composite level (Aim 1). The coverage probability statistic for these methods at the various percentages of missingness are given in Figure 4.49. Multiple imputation produced results at all percentages of missingness that exceed 95%. Coverage using the complete case analysis method dropped below 95% at 40% missingness for all missing data mechanisms except MNAR middle. Median imputation had coverage for all methods except for MAR at 40% missingness. Zero imputation, the recommended method by the creators of the SOFA score, exhibited poor coverage at 20% and greater missingness for the MAR mechanism.

The relative bias statistic for these methods at the various percentages of missingness are given in Figure 4.50 (*MAR* missing data mechanism), Figure 4.51 (*MNAR Left* missing data mechanism), Figure 4.52 (*MNAR Middle* missing data mechanism), and Figure 4.53 (*MNAR Right* missing data mechanism). For the *MAR* missing data mechanism, both median imputation and zero imputation show increasing amounts of bias of the SOFA parameter estimates in the negative direction, however the bias was not enough to affect coverage rates (c.f. Figure 4.49). Both complete case analysis and multiple imputation show relatively unbiased estimates of the SOFA parameter estimate, however the variance of these estimates increases as the percent of missing data increases; moreover, the variance of the relative bias on the multiple imputation estimates is smaller than those of the complete case analysis missing data

method. This pattern of increasing variance with increasing percent of missing data is the same for all missing data mechanisms (c.f. Figures 4.50 through 4.53).

The efficiency statistic for these methods at increasing percentages of missingness are given in Figure 4.54 (*MAR* missing data mechanism), Figure 4.55 (*MNAR Left* missing data mechanism), Figure 4.56 (*MNAR Middle* missing data mechanism), and Figure 4.57 (*MNAR Right* missing data mechanism).

These figures show that with all missing data mechanisms, efficiency rapidly increases—as does the spread of efficiency for the complete case method, showing much larger variance in the SOFA parameter estimates in comparison to the true parameter estimates, and large change in variance across simulation runs (as demonstrated by the spread of these estimates). Similarly, efficiency for the zero imputation method moved away from 1.0 (fully-efficient) as function of increasing percentages of missingness in the negative direction for *MNAR left* (c.f. Figure 4.55) and *MNAR middle* (c.f. Figure 4.56), but in the positive direction for *MNAR right* (c.f. Figure 4.57). Across all of the missing data mechanisms and increasing percentages of missing data, median imputation and MI shows good efficiency—near 1.0—for most of the simulation scenarios, albeit with increasing variance as the percentage of missing data increases.

Figure 4.49 Comparison of percent coverage of the 95% confidence interval for the parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay* (Aim 2 – Component Level)

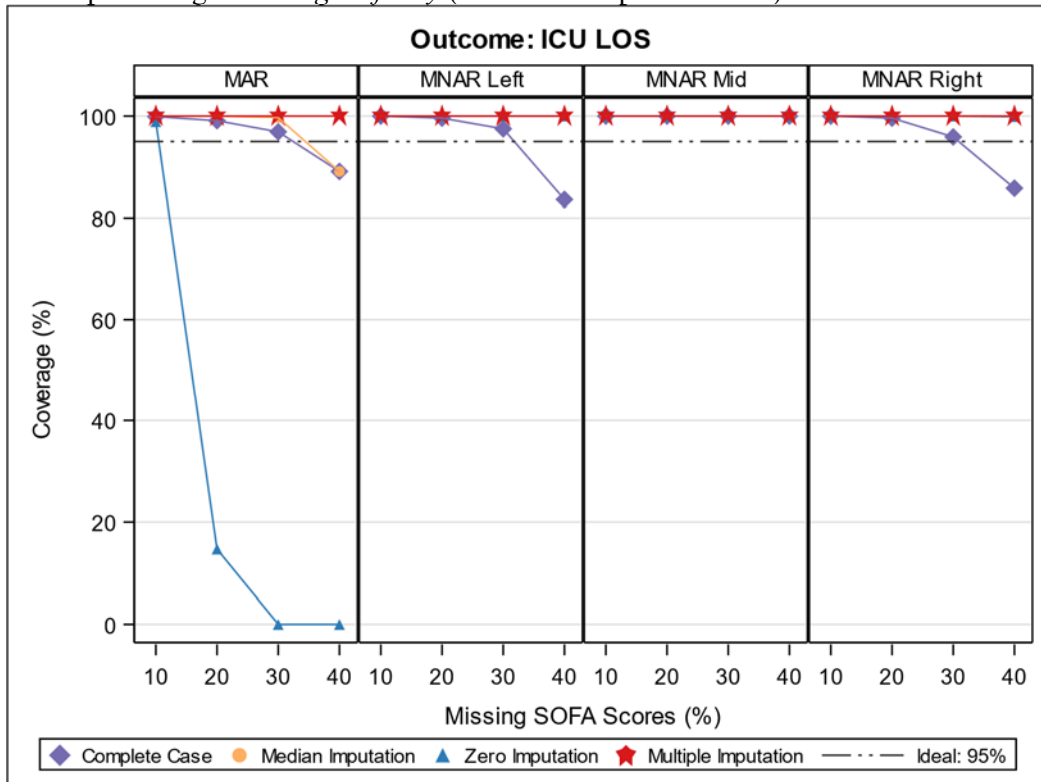


Figure 4.50 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MAR* missing data mechanism (Aim 2 – Component Level)

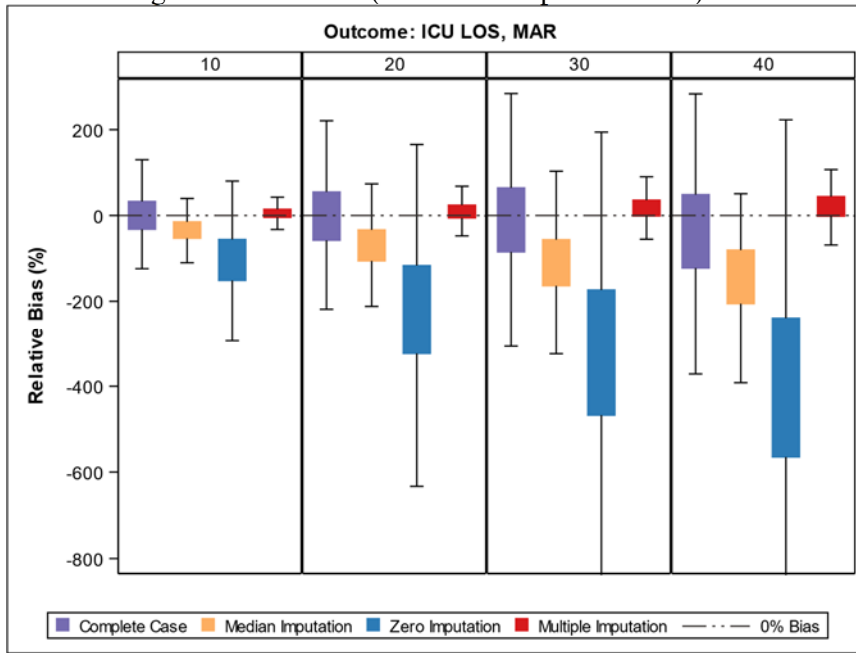


Figure 4.51 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level)

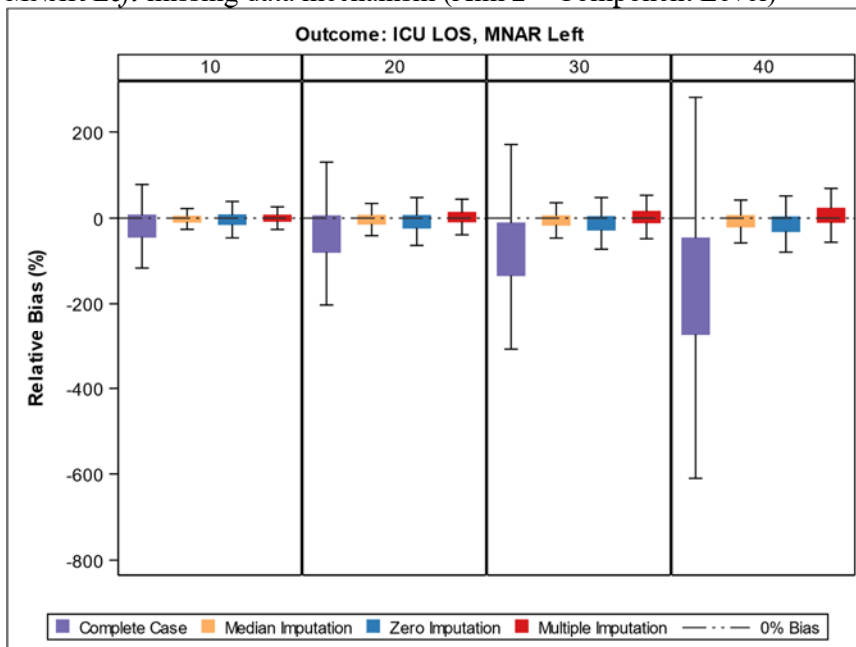


Figure 4.52 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level)

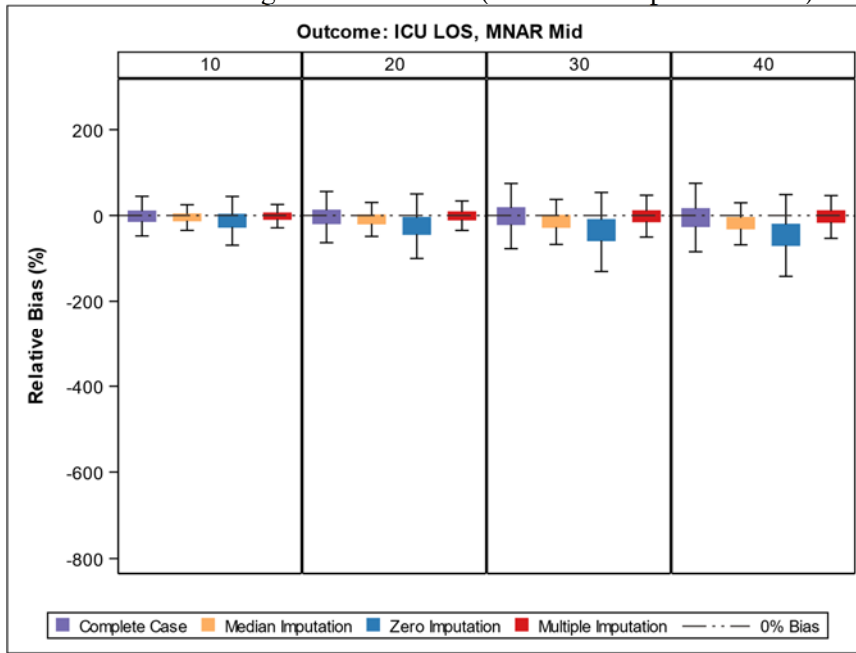


Figure 4.53 Comparison of relative bias of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)

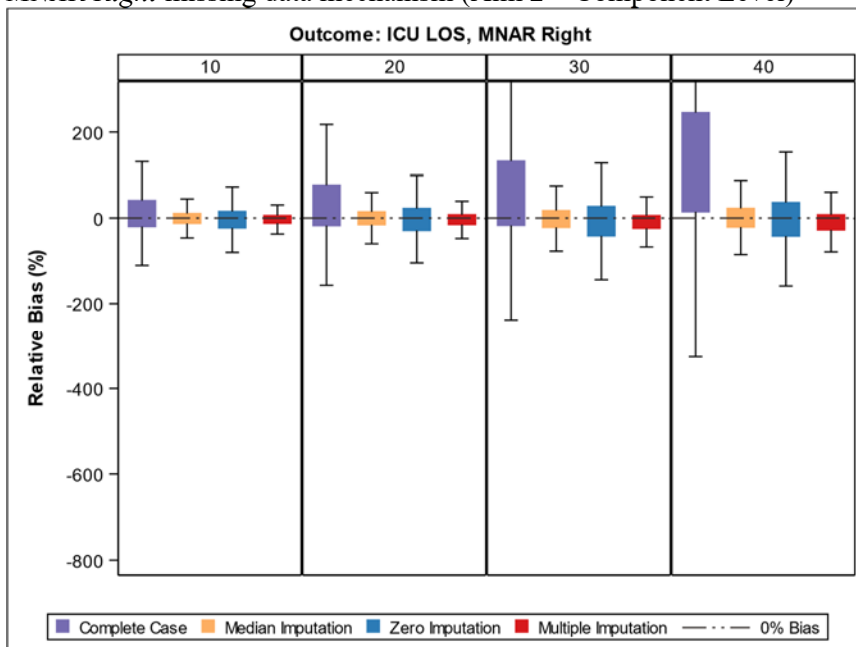


Figure 4.54 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MAR* missing data mechanism (Aim 2 – Component Level)

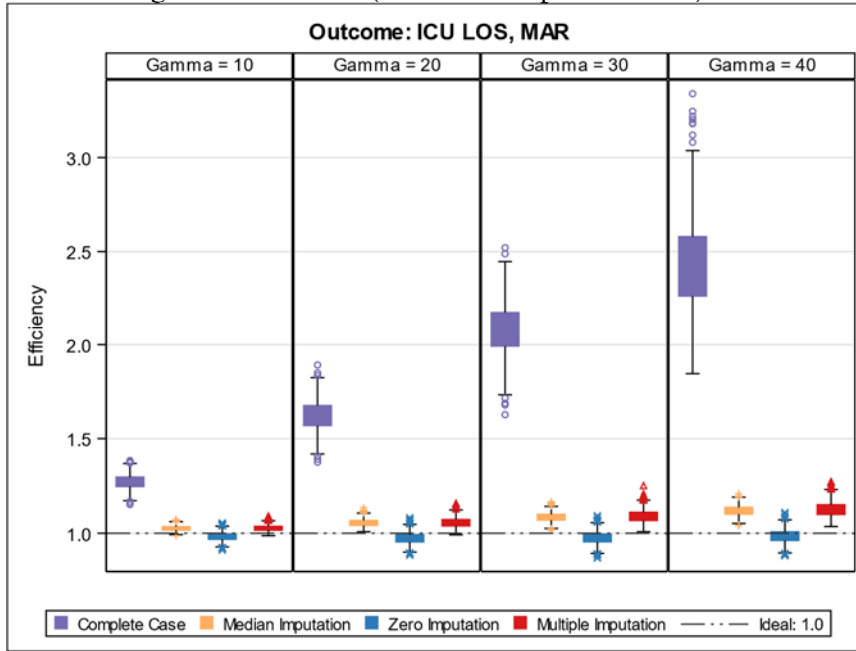


Figure 4.55 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Left* missing data mechanism (Aim 2 – Component Level)

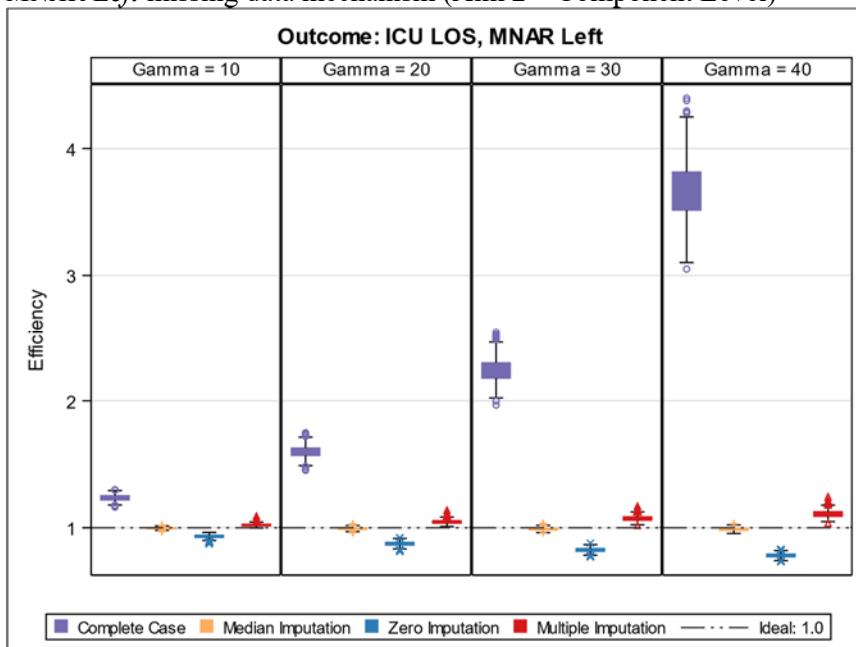


Figure 4.56 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Middle* missing data mechanism (Aim 2 – Component Level)

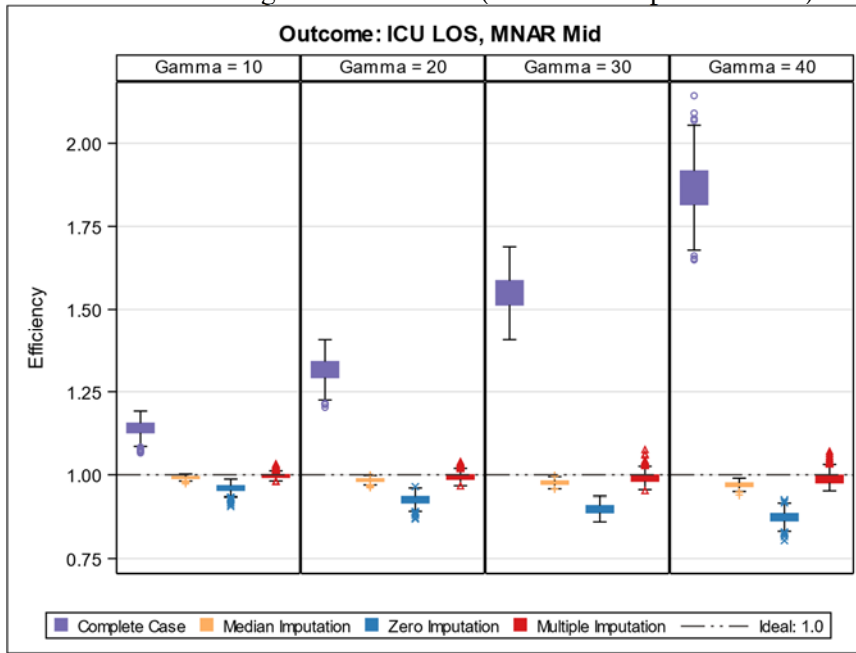
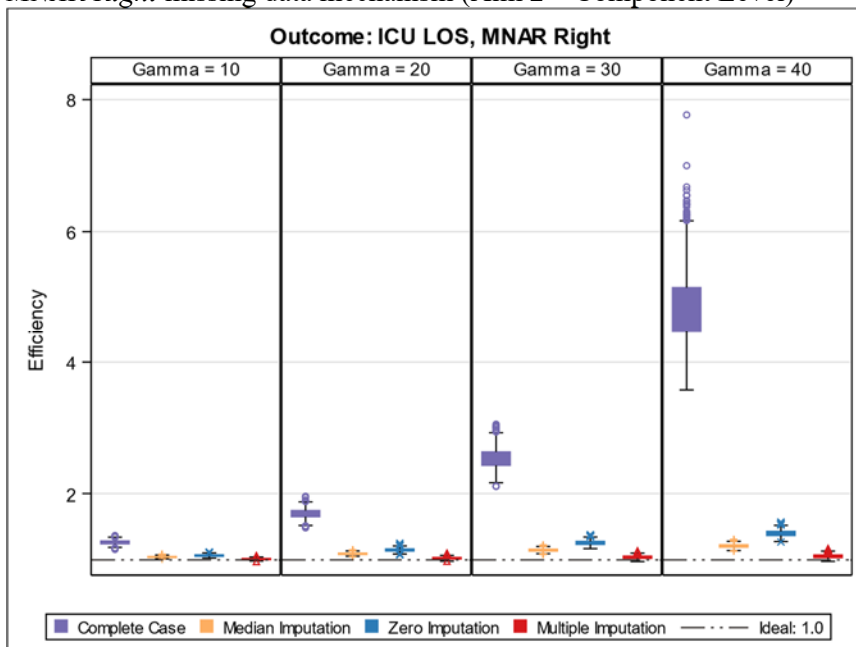


Figure 4.57 Comparison of the efficiency of parameter estimates of the SOFA score in the negative binomial generalized linear model predicting *ICU Length of Stay*, with the *MNAR Right* missing data mechanism (Aim 2 – Component Level)



4.5 Summary and Comparison of Results

This chapter examined the effects of missing SOFA scores at increasing percentages of missingness ($\gamma = 10\%, 20\%, 30\%, 40\%$), under four different missing data mechanism scenarios (MAR, MNAR Left, MNAR Middle, MNAR Right) for three different outcomes (Death, Total charges, ICU LOS), to help researchers understand the effects of methodological choice of method for handling missing data. From the complete dataset ($n=1,930$) simple random sampling from the dataset, with replacement, were made ($n=1,000$). From this dataset parameter estimates were calculated, then missingness was imposed at the given percent of missing data and under the assigned missing data mechanism. The four missing data methods were then applied to this dataset; handling of the missing data was accomplished at the composite level (Aim 1) and the component level (Aim 2). Finally, analyses for the three outcomes were conducted. Results from these analyses of the sampled datasets were compared to those from the complete data.

Overall, for most of the methods studied herein, bias for the SOFA score parameter estimates tended to increase with increasing levels of missingness, as measured by the relative bias statistic. Similarly, the variance of these estimates also tended to increase with increasing levels of missingness, as measured by the efficiency statistic. Finally, coverage probability for the SOFA score tended to decrease with increasing levels of missingness—indicating a higher than expected Type-I error rate. However, *ceteris paribus*, handling missing data at the component level (Aim 2) generally yielded better results than handling missing data at the composite level (Aim 1). Therefore, it is prudent to examine how each of the strategies—composite versus component level—fared by outcome studied.

Regarding the outcome of death in the full dataset, mentioned in Section 4.2.1, the SOFA score was predictive in the adjusted model (OR 1.205, 95% CI 1.174-1.237, $p < 0.0001$). Handling missingness at both the composite and component levels at all percentages of missingness was good for both complete case analysis and multiple imputation methods. Using zero imputation at the composite level (at all percentages of missingness) and the component level (above $\gamma = 20\%$ under MAR; above $\gamma = 30\%$ under MNAR Left and MNAR Middle) results in poor coverage and increasing amounts of bias in the parameter estimates.

Regarding the outcome of total charges in the full dataset, mentioned in Section 4.2.2, the SOFA score was predictive in the adjusted model (β 0.0255, SE 0.0051, $p < 0.0001$). Handling missingness at both the composite and component levels was good for complete case analysis (at or below $\gamma = 30\%$) and multiple imputation (at all percentages of missingness). Using zero imputation at the composite level (above $\gamma = 10\%$ under MNAR Left; at all missingness levels under other missing data mechanisms) and the component level (above $\gamma = 10\%$ under MAR) results in poor coverage and increasing amounts of bias in the parameter estimates.

Regarding the outcome of ICU length of stay in the full dataset, mentioned in Section 4.2.3, the SOFA score was not predictive in the adjusted model (β 0.0081, SE 0.0053, $p = 0.1210$). With that in mind, handling missingness at both the composite and component levels was good for complete case analysis (at or below $\gamma = 30\%$) and multiple imputation (at all percentages of missingness). Using zero imputation at the composite level was poor under all missing data mechanisms, except for MNAR Left. Using zero imputation at the component level worked well under all the MNAR missing data mechanisms studied, and at MAR with $\gamma = 10\%$.

5 DISCUSSION

In December 2015, PCORI convened a workgroup to discuss missing data and data quality for research using electronic medical records and claims data—identifying problems, highlighting current solutions to some of those problems, and suggesting areas for future research [132]. One identified need was for more research to understand the effects of various amounts of missing data, whether the results would be significantly altered by the amount of missing data. Another need was to understand which covariates experience missingness, what the mechanism for missingness is (e.g. MAR), and whether simulations could be used to learn more about these covariates and the impact of missing data. Finally, the PCORI workgroup asserted one of the next steps would be to bring researchers who have experienced success in handling missing data in EMR studies with researchers who are new to EMR studies to help disseminate this knowledge. This study has sought to provide evidence in regards to missingness in the SOFA score within electronic medical records.

5.1 Integration of Findings

The SOFA score is a physiology-based severity of illness score that measures organ derangement in six systems, and is often used in outcomes studies of ICU-treated conditions as a predictor or severity adjustment variable. Severity of illness scores (e.g. SOFA score) differ from comorbidity scores (e.g. Charlson or Elixhauser indices) in that they measure physiological derangement—using clinical and laboratory data—rather than presence or absence of comorbid conditions. As such, severity of illness and comorbidity scores are used in two different manners for risk adjustment, with one adjusting for baseline health (i.e. chronic health) and the other adjusting for severity of illness (i.e. acute health). Therefore, when available, the prudent researcher ought to use physiology-based severity of illness scores in addition to comorbidity scores for baseline risk adjustment. However, when a researcher is unable to calculate a

physiology-based severity of illness score, such as the SOFA score, methodological choices for handling (or ignoring) these missing scores must be made.

This dissertation has examined four techniques for handling missing data for SOFA scores: (1) complete case analysis, (2) median imputation, (3) zero imputation, and (4) Multiple Imputation through Chained Equations (MICE)—a technique that is readily available in the three most common statistical software packages used by applied researchers (SAS, SPSS, and Stata). These techniques have been applied using two approaches, both at the composite level (in support of Aim 1) and at the component level (in support of Aim 2). These techniques were examined using increasing percentages of missingness ($\gamma = 10\%$, 20% , 30% , 40%), under four different missing data mechanism scenarios (MAR, MNAR Left, MNAR Middle, MNAR Right) for three different outcomes (death, total charges, and ICU length of stay), to help researchers understand the effects of methodological choice for handling missing data.

Overall, methods for handling missing data at the component level resulted in superior parameter estimates than handling at the composite level for all methods other than MICE. Multivariate Imputation through Chained Equations, however, had equally good results handling at both the composite and component levels; MICE was demonstrated to be an excellent method for modeling all of the data within the dataset, yielding parameter estimates with excellent coverage, little to no bias, and good efficiency. The value of using an MI approach, of which MICE is but one choice among many, over other approaches tested herein should be discussed.

While in many of the simulations complete case analysis (CCA) resulted in unbiased estimates of the SOFA parameter for many of the outcomes tested, one could see that the precision of the estimate decreased across simulations. Further, the sample size is decreased with this method—yielding lower powered estimates of effect size, and possibly biased estimates of treatment effect and invalid statistical inference due to increased variation. The ramification of a

lower powered study, even when using big data such as in EHR studies, is that the likelihood of committing a Type II error increases—especially when studying a rare outcome.

Median imputation occasionally performed well, however it often resulted in biased estimates and confidence intervals that belied the true certainty around the estimate given the percent of missing data. It is no wonder this method has been rejected by Rubin as being unacceptable for research [113], and has been repeatedly shown in studies to introduce unacceptable bias and over-precise confidence intervals [114], as has been shown again in this study.

The multiple imputation technique of MICE was investigated for its performance. Multiple imputation by chained equations (MICE) demonstrated excellent statistical qualities. In comparison, two of the alternatives tested herein—median and zero imputation, both deterministic imputation techniques—yielded tighter confidence intervals than they should due to lack of accounting for the missing data. Conversely, multiple imputation techniques create multiple datasets which Rubin (1987, p. 2) described as “representing a distribution of possibilities” [98]. As previously stated, the goal of multiple imputation is not to make up data, but rather to allow all the data that are present to be used in analyses to achieve valid statistical inference, not perfect point prediction [113].

Finally, and perhaps most importantly, the research conducted herein directly examined the SOFA score creators’ recommendations on handling missing SOFA score data. According to the Sepsis-3 consensus paper, the baseline SOFA score—which is calculated upon admission to a critical care unit—should be assumed to be zero, unless the patient has a known organ dysfunction [65]. While this may work within the clinical setting, later down the road when research is being conducted such an assertion deserves investigation, as an important methodological choice is being made. In this study, we have demonstrated that zero imputation

results in biased estimates in nearly all of the scenarios examined, yielding estimates that were biased in the negative direction. Because of this, it is the recommendation of this study that zero imputation not be used as a methodological approach for handling missing SOFA scores.

As an alternative, a multiple imputation approach should be considered because of its demonstrated performance; MICE resulted in strong coverage, almost always containing the true parameter estimate, and little bias in most of the scenarios tested in this Monte Carlo simulation study. After proper investigation into potential missing data mechanisms—which will ultimately inform any missing data method—the researcher should opt to use a multiple imputation technique, such as MICE, at either the component or composite level. Consideration should be given to using MI at the composite level, as this can save complexity in statistical programming. Regarding MI methods, MICE in particular should be given special consideration when handling missingness at the component level, as this method does not require the specification of a joint model for all missing variables, but rather as many conditional distributions as there are missing variables. Moreover, when imputing at the component level for a multiple-item instrument, MICE has appealing qualities—such as not requiring the researcher to specify the scale structure, nor the numbers of factors, and does not assume conditional independence of scale items [157].

5.2 Limitations

This research used data from one academic medical center in the Southeastern United States in Charleston, South Carolina. Further, by nature of the study design of requiring a starting point of a complete dataset, the present study was only able to use those observations for which complete SOFA scores could be calculated. Resultantly, at least three limitations arise.

The first limitation is that the overall percentage of missingness in this dataset was 49.1%, with a missing data mechanism believed to be MNAR Left, where lower SOFA scores (those who have lower amounts of organ derangement, and are therefore expected to have a better prognosis,

shorter length of stay, and lower charges) were more likely to be missing. This could have an effect on the interpretation of these results, as patient admissions to the ICU who had little to no organ derangement are under-represented in this study—which leads to another limitation.

The second limitation is in the distribution of the observed SOFA scores within our data. Notwithstanding the aforementioned likely mechanism of higher probability of missing SOFA scores in what is likely the lower end of the SOFA score range, the SOFA scores within our dataset did not span the full range of SOFA scores. The range of SOFA scores in our fully-observed dataset ranged from 0 to 22, whereas the full range of SOFA scores is 0 to 24. This limits generalizability to the very highest SOFA scores, as they were not available in our data. However, it is not known from the literature the approximate percentage of patients within an ICU setting who would be expected to have the two highest SOFA scores (scores of 23 or 24) among patients with ventilator-dependent respiratory failure. If these highest scores are rare, then this limitation would be of less concern.

The third, and final, limitation regards generalizability. This study was performed using data from one academic medical center in the Southeastern United States. Because of this, the performance of the missing data methods may vary based on patient characteristics, although this concern is minimal. This limitation shall be addressed in the following section.

5.3 Future Research

It is the goal of this author to replicate this study using a larger dataset, with ideally a smaller percent of missing data from which to sample for the Monte Carlo simulations. Additionally, is important to test the generalizability of this study's findings, specifically geographic and historical transportability [158], to determine if these findings remain consistent in a different population of patients. The MIMIC-III (Medical Information Mart for Intensive Care) database seems like a viable choice for a replication study, as this database contains over

40,000 patient ICU stays at the Beth Israel Deaconess Medical Center (the academic affiliate of Harvard Medical School) in Boston, Massachusetts between 2001 and 2012 [159]. This source may be an excellent source for a replication study, as it contains rich, longitudinal EHR data including both in- and out-of-hospital mortality on a diverse and large population of ICU patients. Moreover, the data are provided free of charge to researchers worldwide.

Results from this future study, if similar to the results found in the present study, could bolster the argument for the power of multiple imputation methods at the component level for handling large percentages of non-random missing SOFA score data in studies that use electronic health record data—for a variety of research, including quality improvement, comparative effectiveness research, and healthcare cost studies. Importantly, as the setting is in a markedly different geographical region (Northeast United States vs. Southern United States) on a more racially-diverse population, the generalizability of the findings from the current study could be tested.

APPENDICES

Appendix A. Analytical approaches to missing data, search terms

available case analysis	missing
complete case analysis	missing at random
dummy variable adjustment	missing completely at random
entropy balance	missing data
expectation maximization	missing not at random
expectation-maximization	missing value
FCS	missing-at-random
Fully Conditional Specification	missing-completely-at-random
Fully-Conditional Specification	missingness
Heckman	missing-not-at-random
hot deck	MNAR
hot-deck	monotonic
imputation	multiple imputation
incomplete data	NMAR
incomplete observations	non informative missingness
informative missingness	noninformative missingness
inverse probability weighting	non-informative missingness
IPW	nonresponse
last observation carried forward	non-response
LOCF	pattern mixture
Markov Chain Monte Carlo	pattern mixture model
Markov-Chain Monte Carlo	pattern-mixture
MCAR	pattern-mixture model
MCMC	predictive mean matching
mean imputation	selection model
median imputation	

Appendix B. Performance of missing data methods, tables

Table 18 Coverage of the 95% confidence interval for various missing data methods (Aim 1)

<i>Outcome</i>	<i>Mechanism</i>	<i>Gamma</i>	<i>Missing Data Method</i>			
			<i>Complete Case Analysis</i>	<i>Median Imputation</i>	<i>Zero Imputation</i>	<i>Multiple Imputation</i>
Death	MAR	10	100	99.4	0	100
		20	99.9	57.6	0	100
		30	98.8	6.6	0	100
		40	95.4	0	0	99.5
	MNAR Left	10	100	100	33.4	100
		20	99.9	99.9	0	100
		30	100	99.2	0	100
		40	98.4	93.2	0	99.9
	MNAR Middle	10	100	100	1.0	100
		20	100	100	0	100
		30	100	100	0	100
		40	100	100	0	100
	MNAR Right	10	100	100	0	100
		20	100	90.9	0	100
		30	99.7	91.8	0	100
		40	98.9	68.6	0	100
ICU Length of Stay	MAR	10	100	100	0	100
		20	99.2	89.1	0	100
		30	95.7	48.4	0	100
		40	88.0	9.9	0	100
	MNAR Left	10	100	100	99.9	100
		20	99.9	99.6	98.8	100
		30	96.9	97.5	96.0	100
		40	85.1	85.6	96.0	100
	MNAR Middle	10	100	100	91.4	100
		20	100	100	74.0	100
		30	100	100	59.4	100
		40	100	100	42.2	100
	MNAR Right	10	100	100	87.9	100
		20	99.9	99.7	74.9	100
		30	96.2	96.9	68.6	100
		40	87.4	90.7	68.1	100
Total Charges	MAR	10	100	99.5	0	100
		20	98.4	92.3	0	100
		30	96.9	65.7	0	100
		40	90.4	26.4	0	100
	MNAR Left	10	100	100	97.2	100
		20	99.9	99.8	74.3	100
		30	99.0	99.6	49.4	100
		40	95.4	96.2	31.2	100
	MNAR Middle	10	100	100	68.3	100
		20	100	100	23.3	100
		30	100	100	11.3	100
		40	100	100	4.6	100
	MNAR Right	10	100	100	49.7	100
		20	99.7	98.6	21.0	100
		30	98.1	97.7	13.1	100
		40	94.4	95.0	6.0	100

Table 19 Coverage of the 95% confidence interval for various missing data methods (Aim 2)

<i>Outcome</i>	<i>Mechanism</i>	<i>Gamma</i>	<i>Missing Data Method</i>			
			<i>Complete Case Analysis</i>	<i>Median Imputation</i>	<i>Zero Imputation</i>	<i>Multiple Imputation</i>
Death	MAR	10	100	100	100	100
		20	100	100	2.4	100
		30	98.0	99.9	0	100
		40	95.4	93.5	0	100
	MNAR Left	10	100	100	100	100
		20	100	100	100	100
		30	100	100	99.3	100
		40	99.3	100	86.4	100
	MNAR Middle	10	100	100	100	100
		20	100	100	100	100
		30	100	100	99.2	100
		40	100	100	93.8	100
	MNAR Right	10	100	100	100	100
		20	100	100	100	100
		30	99.8	100	100	100
		40	99.0	100	100	100
ICU Length of Stay	MAR	10	99.9	100	98.9	100
		20	99.1	100	14.7	100
		30	96.9	99.7	0	100
		40	89.1	89.1	0	100
	MNAR Left	10	100	100	100	100
		20	99.6	100	100	100
		30	97.5	100	100	100
		40	83.6	100	100	100
	MNAR Middle	10	100	100	100	100
		20	100	100	100	100
		30	100	100	100	100
		40	100	100	100	100
	MNAR Right	10	100	100	100	100
		20	99.6	100	100	100
		30	95.9	100	100	100
		40	85.8	100	99.9	100
Total Charges	MAR	10	99.9	100	99.6	100
		20	99.6	100	38.6	100
		30	96.0	100	0.3	100
		40	90.8	99.1	0	100
	MNAR Left	10	100	100	100	100
		20	99.8	100	100	100
		30	99.3	100	100	100
		40	95.6	100	100	100
	MNAR Middle	10	100	100	100	100
		20	100	100	100	100
		30	100	100	100	100
		40	100	100	99.9	100
	MNAR Right	10	100	100	100	100
		20	99.7	100	100	100
		30	97.6	100	99.9	100
		40	92.5	100	100	100

Table 20 Relative bias for various missing data methods (Aim 1)

<i>Outcome</i>	<i>Mechanism</i>	<i>Gamma</i>	<i>Missing Data Method</i>			
			<i>Complete Case Analysis</i>	<i>Median Imputation</i>	<i>Zero Imputation</i>	<i>Multiple Imputation</i>
Death	MAR	10	0.011 (0.265)	-0.029 (0.108)	-0.061 (0.372)	-0.001 (0.045)
		20	0.027 (0.427)	-0.050 (0.178)	-0.086 (0.521)	-0.002 (0.030)
		30	0.053 (0.657)	-0.071 (0.242)	-0.141 (0.585)	0.006 (0.042)
		40	0.084 (1.106)	-0.079 (0.447)	-0.248 (0.586)	0.007 (0.050)
	MNAR Left	10	-0.035 (0.187)	0.002 (0.070)	-0.016 (0.093)	-2.753 (0.027)
		20	-0.082 (0.298)	0.002 (0.127)	-0.025 (0.119)	0.000 (0.016)
		30	-0.144 (0.407)	0.014 (0.170)	-0.029 (0.136)	-0.000 (0.017)
		40	-0.221 (0.564)	0.021 (0.251)	-0.028 (0.134)	-0.000 (0.022)
	MNAR Middle	10	0.004 (0.212)	-0.002 (0.031)	-0.030 (0.167)	0.001 (0.041)
		20	0.014 (0.306)	-0.004 (0.042)	-0.041 (0.229)	0.003 (0.024)
		30	0.027 (0.404)	-0.005 (0.049)	-0.046 (0.265)	-0.002 (0.029)
		40	0.043 (0.507)	-0.009 (0.055)	-0.050 (0.289)	-0.002 (0.034)
	MNAR Right	10	0.034 (0.232)	-0.013 (0.105)	-0.051 (0.297)	-0.005 (0.067)
		20	0.092 (0.366)	-0.030 (0.186)	-0.061 (0.408)	-0.006 (0.050)
		30	0.171 (0.528)	-0.030 (0.221)	-0.072 (0.456)	0.014 (0.064)
		40	0.300 (0.773)	-0.052 (0.314)	-0.079 (0.462)	0.016 (0.075)
ICU Length of Stay	MAR	10	0.013 (0.802)	0.002 (0.060)	0.004 (0.372)	-0.000 (0.041)
		20	-0.022 (1.375)	0.002 (0.107)	-0.038 (0.686)	0.001 (0.026)
		30	-0.119 (1.943)	0.003 (0.163)	-0.087 (1.010)	-0.000 (0.025)
		40	-0.384 (2.586)	-0.001 (0.244)	-0.220 (1.305)	-0.001 (0.032)
	MNAR Left	10	0.010 (0.687)	0.001 (0.038)	-0.000 (0.042)	0.002 (0.061)
		20	-0.012 (1.156)	0.001 (0.066)	-0.000 (0.053)	0.006 (0.038)
		30	0.020 (1.549)	0.003 (0.079)	-0.000 (0.052)	-0.005 (0.044)
		40	0.027 (2.104)	0.003 (0.110)	-0.000 (0.051)	-0.003 (0.054)
	MNAR Middle	10	-0.003 (0.570)	0.000 (0.016)	0.000 (0.070)	-0.008 (0.090)
		20	0.005 (0.872)	0.001 (0.025)	0.001 (0.091)	-0.011 (0.070)
		30	0.017 (1.145)	0.002 (0.027)	0.002 (0.099)	0.025 (0.086)
		40	0.024 (1.433)	0.002 (0.033)	-0.000 (0.128)	0.030 (0.094)
	MNAR Right	10	-0.019 (0.654)	0.000 (0.047)	0.001 (0.097)	-0.000 (0.053)
		20	-0.072 (1.086)	0.001 (0.073)	0.001 (0.114)	0.002 (0.035)
		30	-0.080 (1.527)	0.000 (0.091)	0.000 (0.119)	-0.002 (0.034)
		40	-0.156 (2.079)	0.004 (0.120)	0.002 (0.117)	-0.002 (0.043)
Total Charges	MAR	10	-0.006 (0.437)	0.009 (0.076)	0.033 (0.243)	0.005 (0.076)
		20	-0.029 (0.717)	0.017 (0.120)	0.039 (0.425)	0.011 (0.049)
		30	-0.079 (0.948)	0.024 (0.158)	0.030 (0.621)	-0.010 (0.057)
		40	-0.195 (1.215)	0.027 (0.187)	-0.017 (0.801)	-0.006 (0.068)
	MNAR Left	10	0.058 (0.365)	0.006 (0.046)	0.005 (0.074)	-0.013 (0.113)
		20	0.155 (0.591)	0.013 (0.076)	0.009 (0.088)	-0.018 (0.103)
		30	0.273 (0.841)	0.019 (0.089)	0.012 (0.095)	0.041 (0.112)
		40	0.454 (1.138)	0.026 (0.113)	0.014 (0.093)	0.051 (0.116)
	MNAR Middle	10	-0.023 (0.342)	0.002 (0.027)	0.013 (0.108)	-0.000 (0.062)
		20	-0.053 (0.527)	0.005 (0.039)	0.022 (0.137)	0.003 (0.045)
		30	-0.121 (0.695)	0.008 (0.046)	0.027 (0.137)	-0.003 (0.045)
		40	-0.174 (0.866)	0.011 (0.053)	0.030 (0.139)	-0.003 (0.053)
	MNAR Right	10	-0.052 (0.365)	0.002 (0.072)	0.012 (0.164)	0.007 (0.090)
		20	-0.158 (0.597)	0.007 (0.116)	0.018 (0.188)	0.017 (0.063)
		30	-0.252 (0.817)	0.008 (0.133)	0.024 (0.195)	-0.017 (0.069)
		40	-0.429 (1.088)	0.018 (0.165)	0.029 (0.183)	-0.011 (0.093)

Table 21 Relative bias for various missing data methods (Aim 2)

<i>Outcome</i>	<i>Mechanism</i>	<i>Gamma</i>	<i>Missing Data Method</i>			
			<i>Complete Case Analysis</i>	<i>Median Imputation</i>	<i>Zero Imputation</i>	<i>Multiple Imputation</i>
Death	MAR	10	0.009 (0.258)	-0.008 (0.040)	-0.012 (0.084)	-0.001 (0.025)
		20	0.028 (0.420)	-0.015 (0.066)	-0.023 (0.141)	-0.002 (0.037)
		30	0.050 (0.658)	-0.022 (0.085)	-0.032 (0.199)	-0.004 (0.049)
		40	0.060 (1.085)	-0.027 (0.109)	-0.038 (0.254)	-0.004 (0.060)
	MNAR Left	10	-0.034 (0.183)	-0.002 (0.016)	-0.005 (0.033)	0.001 (0.018)
		20	-0.085 (0.291)	-0.004 (0.025)	-0.009 (0.046)	0.003 (0.027)
		30	-0.133 (0.406)	-0.006 (0.032)	-0.012 (0.056)	0.005 (0.035)
		40	-0.219 (0.552)	-0.008 (0.037)	-0.015 (0.061)	0.007 (0.042)
	MNAR Middle	10	0.004 (0.213)	-0.004 (0.024)	-0.007 (0.045)	-0.001 (0.023)
		20	0.013 (0.312)	-0.008 (0.035)	-0.014 (0.067)	-0.002 (0.034)
		30	0.022 (0.410)	-0.011 (0.044)	-0.018 (0.082)	-0.002 (0.043)
		40	0.040 (0.490)	-0.014 (0.051)	-0.024 (0.094)	-0.003 (0.051)
	MNAR Right	10	0.032 (0.229)	-0.001 (0.035)	-0.002 (0.055)	-0.001 (0.026)
		20	0.097 (0.372)	-0.004 (0.050)	-0.005 (0.077)	-0.002 (0.039)
		30	0.163 (0.548)	-0.007 (0.061)	-0.006 (0.091)	-0.005 (0.048)
		40	0.319 (0.757)	-0.008 (0.072)	-0.004 (0.109)	-0.007 (0.057)
ICU Length of Stay	MAR	10	-0.000 (0.812)	0.001 (0.031)	0.004 (0.087)	-0.000 (0.015)
		20	-0.018 (1.418)	0.003 (0.060)	0.008 (0.181)	-0.000 (0.023)
		30	-0.139 (1.949)	0.003 (0.092)	0.005 (0.270)	-0.001 (0.031)
		40	-0.373 (2.621)	0.004 (0.123)	0.002 (0.350)	-0.001 (0.038)
	MNAR Left	10	0.006 (0.664)	0.000 (0.009)	-0.000 (0.015)	0.000 (0.010)
		20	0.016 (1.125)	0.000 (0.013)	-0.000 (0.021)	9.358 (0.014)
		30	-0.001 (1.558)	0.000 (0.017)	-0.000 (0.026)	-6.661 (0.019)
		40	0.033 (2.073)	0.000 (0.019)	-0.000 (0.027)	-0.000 (0.021)
	MNAR Middle	10	-0.002 (0.552)	0.000 (0.011)	0.000 (0.021)	3.916 (0.010)
		20	-0.006 (0.870)	0.000 (0.017)	0.000 (0.033)	-8.288 (0.016)
		30	0.020 (1.122)	0.000 (0.020)	0.001 (0.037)	0.000 (0.018)
		40	0.001 (1.412)	0.001 (0.024)	0.001 (0.048)	-0.000 (0.022)
	MNAR Right	10	-0.019 (0.651)	0.000 (0.015)	0.000 (0.025)	6.951 (0.011)
		20	-0.064 (1.042)	0.000 (0.022)	0.000 (0.036)	-0.000 (0.018)
		30	-0.094 (1.536)	0.000 (0.027)	0.001 (0.044)	-0.000 (0.023)
		40	-0.139 (2.137)	0.000 (0.029)	0.001 (0.049)	-5.602 (0.028)
Total Charges	MAR	10	-0.006 (0.452)	0.003 (0.035)	0.009 (0.070)	0.000 (0.023)
		20	-0.039 (0.702)	0.007 (0.053)	0.019 (0.114)	0.001 (0.034)
		30	-0.075 (0.962)	0.010 (0.071)	0.025 (0.163)	0.001 (0.042)
		40	-0.191 (1.218)	0.013 (0.091)	0.032 (0.211)	0.000 (0.052)
	MNAR Left	10	0.058 (0.356)	0.000 (0.013)	-0.000 (0.024)	0.000 (0.015)
		20	0.158 (0.593)	0.001 (0.019)	0.000 (0.035)	0.000 (0.022)
		30	0.290 (0.851)	0.001 (0.024)	0.000 (0.041)	0.001 (0.028)
		40	0.456 (1.129)	0.002 (0.030)	0.001 (0.045)	0.001 (0.035)
	MNAR Middle	10	-0.021 (0.335)	0.000 (0.018)	0.000 (0.034)	-0.000 (0.017)
		20	-0.065 (0.539)	0.001 (0.027)	0.004 (0.049)	-0.001 (0.026)
		30	-0.099 (0.702)	0.001 (0.032)	0.005 (0.056)	-0.002 (0.032)
		40	-0.187 (0.883)	0.002 (0.038)	0.008 (0.063)	-0.003 (0.037)
	MNAR Right	10	-0.060 (0.379)	0.001 (0.025)	0.001 (0.041)	0.000 (0.019)
		20	-0.153 (0.601)	0.002 (0.037)	0.003 (0.060)	0.001 (0.029)
		30	-0.274 (0.850)	0.004 (0.043)	0.006 (0.069)	0.002 (0.037)
		40	-0.419 (1.106)	0.006 (0.049)	0.008 (0.078)	0.003 (0.045)

Table 22 Efficiency for various missing data methods (Aim 1)

<i>Outcome</i>	<i>Mechanism</i>	<i>Gamma</i>	<i>Missing Data Method</i>			
			<i>Complete Case Analysis</i>	<i>Median Imputation</i>	<i>Zero Imputation</i>	<i>Multiple Imputation</i>
Death	MAR	10	1.315 (0.070)	0.971 (0.014)	0.894 (0.023)	1.004 (0.010)
		20	1.856 (0.178)	0.946 (0.019)	0.889 (0.031)	1.009 (0.016)
		30	3.006 (0.516)	0.923 (0.023)	0.935 (0.037)	1.016 (0.024)
		40	6.860 (2.641)	0.890 (0.028)	1.082 (0.083)	1.025 (0.039)
	MNAR Left	10	1.212 (0.043)	0.991 (0.011)	0.980 (0.014)	1.003 (0.006)
		20	1.505 (0.085)	0.975 (0.017)	0.972 (0.016)	1.008 (0.010)
		30	1.923 (0.144)	0.969 (0.019)	0.970 (0.016)	1.013 (0.014)
		40	2.557 (0.242)	0.950 (0.023)	0.971 (0.016)	1.021 (0.021)
	MNAR Middle	10	1.242 (0.059)	0.998 (0.005)	0.963 (0.019)	1.004 (0.007)
		20	1.590 (0.110)	0.997 (0.006)	0.950 (0.022)	1.010 (0.011)
		30	2.115 (0.184)	0.996 (0.008)	0.943 (0.021)	1.016 (0.015)
		40	2.945 (0.288)	0.994 (0.008)	0.942 (0.021)	1.023 (0.019)
	MNAR Right	10	1.264 (0.057)	0.981 (0.016)	0.922 (0.024)	1.004 (0.010)
		20	1.660 (0.133)	0.955 (0.023)	0.896 (0.028)	1.009 (0.015)
		30	2.299 (0.326)	0.941 (0.026)	0.885 (0.028)	1.014 (0.020)
		40	3.470 (0.982)	0.913 (0.027)	0.890 (0.027)	1.017 (0.024)
ICU Length of Stay	MAR	10	1.271 (0.045)	1.000 (0.002)	0.989 (0.012)	1.000 (0.001)
		20	1.641 (0.095)	1.001 (0.004)	0.954 (0.021)	1.000 (0.002)
		30	2.112 (0.172)	1.000 (0.005)	0.888 (0.037)	1.001 (0.003)
		40	2.531 (0.318)	0.997 (0.007)	0.769 (0.052)	1.002 (0.004)
	MNAR Left	10	1.243 (0.038)	1.000 (0.001)	0.999 (0.002)	1.000 (0.000)
		20	1.595 (0.076)	1.001 (0.002)	1.000 (0.002)	1.000 (0.001)
		30	2.124 (0.159)	1.001 (0.003)	1.000 (0.002)	1.000 (0.002)
		40	2.947 (0.321)	1.001 (0.004)	1.000 (0.002)	1.001 (0.003)
	MNAR Middle	10	1.242 (0.046)	1.000 (0.000)	1.000 (0.003)	1.000 (0.000)
		20	1.590 (0.088)	1.000 (0.001)	1.000 (0.003)	1.000 (0.001)
		30	2.111 (0.141)	1.000 (0.001)	1.000 (0.004)	1.000 (0.001)
		40	2.925 (0.232)	1.000 (0.001)	1.000 (0.004)	1.000 (0.002)
	MNAR Right	10	1.225 (0.042)	1.000 (0.002)	0.999 (0.004)	1.000 (0.001)
		20	1.531 (0.087)	1.000 (0.003)	1.000 (0.005)	1.000 (0.001)
		30	1.970 (0.150)	0.999 (0.004)	1.000 (0.005)	1.000 (0.002)
		40	2.617 (0.253)	1.000 (0.005)	1.000 (0.005)	1.000 (0.002)
Total Charges	MAR	10	1.251 (0.045)	1.004 (0.005)	1.008 (0.012)	1.001 (0.003)
		20	1.593 (0.087)	1.007 (0.007)	0.991 (0.019)	1.003 (0.005)
		30	2.031 (0.153)	1.009 (0.008)	0.951 (0.032)	1.005 (0.006)
		40	2.485 (0.259)	1.010 (0.009)	0.879 (0.041)	1.008 (0.008)
	MNAR Left	10	1.242 (0.041)	1.001 (0.003)	1.000 (0.004)	1.000 (0.002)
		20	1.592 (0.080)	1.003 (0.005)	1.001 (0.005)	1.001 (0.003)
		30	2.121 (0.162)	1.005 (0.006)	1.001 (0.005)	1.003 (0.004)
		40	2.940 (0.340)	1.007 (0.008)	1.001 (0.004)	1.005 (0.006)
	MNAR Middle	10	1.241 (0.044)	1.000 (0.002)	1.003 (0.007)	1.000 (0.002)
		20	1.587 (0.084)	1.001 (0.002)	1.005 (0.007)	1.001 (0.003)
		30	2.104 (0.134)	1.001 (0.003)	1.006 (0.008)	1.002 (0.004)
		40	2.916 (0.224)	1.002 (0.003)	1.007 (0.008)	1.003 (0.005)
	MNAR Right	10	1.226 (0.042)	1.001 (0.005)	1.005 (0.010)	1.000 (0.002)
		20	1.535 (0.086)	1.003 (0.008)	1.007 (0.010)	1.001 (0.004)
		30	1.978 (0.145)	1.004 (0.009)	1.008 (0.011)	1.002 (0.005)
		40	2.637 (0.257)	1.007 (0.010)	1.008 (0.010)	1.002 (0.006)

Table 23 Efficiency for various missing data methods (Aim 2)

<i>Outcome</i>	<i>Mechanism</i>	<i>Gamma</i>	<i>Missing Data Method</i>			
			<i>Complete Case Analysis</i>	<i>Median Imputation</i>	<i>Zero Imputation</i>	<i>Multiple Imputation</i>
Death	MAR	10	1.314 (0.069)	1.272 (0.047)	1.251 (0.045)	0.989 (0.006)
		20	1.000 (0.001)	1.002 (0.002)	0.971 (0.008)	1.000 (0.002)
		30	1.004 (0.003)	1.002 (0.005)	1.000 (0.000)	1.000 (0.001)
		40	1.857 (0.173)	1.638 (0.095)	1.589 (0.088)	0.978 (0.008)
	MNAR Left	10	1.001 (0.002)	1.003 (0.003)	0.947 (0.012)	0.999 (0.005)
		20	1.008 (0.005)	1.004 (0.008)	1.000 (0.001)	1.000 (0.002)
		30	3.003 (0.531)	2.110 (0.175)	2.029 (0.156)	0.969 (0.010)
		40	1.001 (0.002)	1.005 (0.003)	0.929 (0.014)	0.996 (0.007)
	MNAR Middle	10	1.010 (0.007)	1.007 (0.011)	1.000 (0.001)	1.001 (0.003)
		20	6.811 (2.671)	2.534 (0.315)	2.491 (0.254)	0.960 (0.012)
		30	1.001 (0.003)	1.007 (0.004)	0.916 (0.017)	0.991 (0.009)
		40	1.010 (0.009)	1.009 (0.013)	1.000 (0.001)	1.001 (0.003)
	MNAR Right	10	1.212 (0.043)	1.244 (0.038)	1.244 (0.041)	0.998 (0.003)
		20	1.000 (0.000)	0.999 (0.000)	0.995 (0.006)	0.999 (0.000)
		30	0.999 (0.001)	1.001 (0.003)	1.000 (0.000)	1.000 (0.001)
		40	1.504 (0.086)	1.592 (0.075)	1.590 (0.081)	0.996 (0.004)
ICU Length of Stay	MAR	10	1.000 (0.000)	0.999 (0.001)	0.992 (0.007)	0.999 (0.001)
		20	0.999 (0.002)	1.002 (0.005)	1.000 (0.000)	1.000 (0.001)
		30	1.924 (0.151)	2.122 (0.152)	2.116 (0.158)	0.995 (0.006)
		40	1.000 (0.000)	0.999 (0.001)	0.990 (0.009)	0.999 (0.001)
	MNAR Left	10	0.999 (0.002)	1.003 (0.006)	1.000 (0.001)	1.000 (0.002)
		20	2.547 (0.231)	2.957 (0.319)	2.950 (0.338)	0.993 (0.006)
		30	1.000 (0.000)	0.999 (0.001)	0.988 (0.010)	0.999 (0.001)
		40	0.999 (0.002)	1.005 (0.007)	1.000 (0.001)	1.000 (0.002)
	MNAR Middle	10	1.241 (0.059)	1.242 (0.046)	1.242 (0.044)	0.997 (0.003)
		20	1.000 (0.000)	1.000 (0.001)	0.994 (0.007)	1.000 (0.000)
		30	1.000 (0.002)	1.000 (0.003)	1.000 (0.000)	1.000 (0.001)
		40	1.591 (0.102)	1.589 (0.086)	1.589 (0.081)	0.995 (0.005)
	MNAR Right	10	1.000 (0.000)	1.000 (0.001)	0.990 (0.010)	1.000 (0.001)
		20	1.000 (0.003)	1.000 (0.005)	1.000 (0.000)	1.000 (0.001)
		30	2.111 (0.179)	2.108 (0.147)	2.107 (0.141)	0.993 (0.006)
		40	1.000 (0.000)	1.000 (0.002)	0.987 (0.011)	1.000 (0.001)
Total Charges	MAR	10	1.001 (0.003)	1.000 (0.006)	1.000 (0.000)	1.000 (0.002)
		20	2.943 (0.293)	2.927 (0.231)	2.916 (0.221)	0.991 (0.007)
		30	1.000 (0.001)	1.000 (0.002)	0.985 (0.012)	1.000 (0.001)
		40	1.001 (0.003)	1.000 (0.008)	1.000 (0.001)	1.001 (0.002)
	MNAR Left	10	1.263 (0.056)	1.225 (0.043)	1.227 (0.042)	0.997 (0.006)
		20	1.000 (0.000)	1.000 (0.002)	0.994 (0.010)	1.000 (0.001)
		30	1.000 (0.003)	1.000 (0.005)	1.000 (0.000)	1.000 (0.001)
		40	1.661 (0.141)	1.535 (0.085)	1.535 (0.085)	0.993 (0.009)
	MNAR Middle	10	1.000 (0.001)	1.000 (0.002)	0.987 (0.014)	1.000 (0.001)
		20	1.001 (0.004)	1.001 (0.007)	1.000 (0.000)	1.001 (0.002)
		30	2.300 (0.335)	1.969 (0.150)	1.975 (0.147)	0.989 (0.010)
		40	1.000 (0.001)	1.001 (0.003)	0.982 (0.015)	1.000 (0.002)
	MNAR Right	10	1.002 (0.005)	1.000 (0.009)	1.000 (0.001)	1.002 (0.002)
		20	3.480 (0.935)	2.619 (0.252)	2.640 (0.254)	0.986 (0.012)
		30	1.000 (0.001)	1.001 (0.003)	0.978 (0.017)	1.000 (0.002)
		40	1.003 (0.005)	1.001 (0.011)	1.000 (0.001)	1.002 (0.003)

Appendix C. Correlation table

Table 24 Intercorrelations for variables used in this study, measured by Spearman's rank correlation coefficient, ρ_s

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 SOFA	—	0.233 <i><.0001</i>	0.698 <i><.0001</i>	0.484 <i><.0001</i>	0.467 <i><.0001</i>	0.560 <i><.0001</i>	0.605 <i><.0001</i>	0.014 <i>0.534</i>	0.048 0.034	-0.038 0.093	0.311 <i><.0001</i>	0.088 0.0001	0.035 <i>0.120</i>	0.229 <i><.0001</i>
2 SOFA CNS	—	—	-0.014 <i>0.534</i>	-0.045 0.046	-0.050 0.029	-0.003 <i>0.881</i>	-0.060 0.008	-0.024 <i>0.294</i>	-0.018 <i>0.419</i>	-0.023 <i>0.322</i>	0.150 <i><.0001</i>	-0.089 <i><.0001</i>	-0.037 <i>0.103</i>	-0.053 0.019
3 SOFA Card	—	—	—	0.204 <i><.0001</i>	0.218 <i><.0001</i>	0.267 <i><.0001</i>	0.349 <i><.0001</i>	0.045 0.046	0.001 <i>0.965</i>	0.053 0.019	0.232 <i><.0001</i>	-0.003 <i>0.892</i>	-0.046 0.043	0.067 0.003
4 SOFA Coag	—	—	—	—	0.364 <i><.0001</i>	0.158 <i><.0001</i>	0.169 <i><.0001</i>	-0.017 <i>0.445</i>	0.042 0.065	0.025 <i>0.268</i>	0.155 <i><.0001</i>	0.166 <i><.0001</i>	0.083 0.001	0.154 <i><.0001</i>
5 SOFA Hep	—	—	—	—	—	0.201 <i><.0001</i>	0.131 <i><.0001</i>	-0.019 <i>0.410</i>	0.068 0.003	0.012 <i>0.613</i>	0.189 <i><.0001</i>	0.124 <i><.0001</i>	0.075 0.001	0.171 <i><.0001</i>
6 SOFA Renal	—	—	—	—	—	—	0.177 <i><.0001</i>	0.065 0.004	0.073 0.001	-0.186 <i><.0001</i>	0.158 <i><.0001</i>	-0.014 <i>0.525</i>	-0.023 0.321	0.313 <i><.0001</i>
7 SOFA Resp	—	—	—	—	—	—	—	-0.012 <i>0.588</i>	0.059 0.009	0.021 <i>0.345</i>	0.137 <i><.0001</i>	0.126 <i><.0001</i>	0.077 0.001	0.074 0.001
8 Age Group	—	—	—	—	—	—	—	—	-0.025 <i>0.268</i>	0.053 0.020	0.168 <i><.0001</i>	-0.087 0.0001	-0.018 <i>0.418</i>	0.253 <i><.0001</i>
9 Male	—	—	—	—	—	—	—	—	—	0.069 0.002	-0.056 0.014	0.046 0.044	0.017 <i>0.449</i>	-0.033 <i>0.149</i>
10 Race	—	—	—	—	—	—	—	—	—	—	0.057 0.013	-0.030 0.194	-0.044 0.053	-0.104 <i><.0001</i>
11 Died	—	—	—	—	—	—	—	—	—	—	—	-0.171 <i><.0001</i>	-0.169 <i><.0001</i>	0.134 <i><.0001</i>
12 Total Charges	—	—	—	—	—	—	—	—	—	—	—	—	0.813 <i><.0001</i>	0.155 <i><.0001</i>
13 ICU LOS	—	—	—	—	—	—	—	—	—	—	—	—	—	0.152 <i><.0001</i>
14 Charlson Score	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Italicized indicates the p-value: prob > |r| under $H_0: \rho_s=0$

Bold indicates $p < 0.05$

Appendix D. Example SAS Code

Missing data generation macro – MAR missing data mechanism

This macro selects records based on the MAR missing data mechanism, whereby 75% of the $\gamma\%$ of observations are selected at random among patients who died during their admission and 25% of the $\gamma\%$ of observations are selected at random among patients whose Charlson comorbidity score is ≥ 2 and who survived the admission. A Charlson comorbidity score of 2 is the median value among patients who survived the admission.

```
%macro mar(in=,out=,percent=);
  data _NULL_; *** output the # obs to variable nrows;
    if 0 then set &in nobs=n;
    call symputx('nrows',n);
    stop;
  run;
  %let n_SetMissing = %sysevalf(&nrows *
                              (&percent/100),ceil);
  *round up to the nearest integer (e.g. 3.1 becomes 4);
  %let n_SetMissing1 = %sysevalf(&n_SetMissing *
                                0.75,ceil);
  *75% from Died=1;

  %let n_SetMissing2 = %sysevalf(&n_SetMissing -
                                &n_SetMissing1,ceil);
  *25% from Died=0 and CharlsScore > 2;

  data _temp1;
    set &in;
    where died=1;
    probb_SOFA_miss = rand("normal",0,1);
  run;

  data _temp2;
    set &in;
    where died=0;
    if CharlsScore > 2 then probb_SOFA_miss =
      rand("normal",0,1);
      else probb_SOFA_miss = -9;
  run;

  proc sort data=_temp1;
    by descending probb_SOFA_miss;
  run;
```

```
data _temp1;
  set _temp1;
  if _n_ <= &n_SetMissing1 then Selected=1;
  else Selected=0;
  drop prob_SOFA_miss;
run;

proc sort data=_temp2;
  by descending prob_SOFA_miss;
run;

data _temp2;
  set _temp2;
  if _n_ <= &n_SetMissing2 then Selected=1;
  else Selected=0;
  drop prob_SOFA_miss;
run;

data &out;
  set _temp1 _temp2;
run;

proc delete data=_temp1; run;
proc delete data=_temp2; run;

%mend;
```

Missing data generation macro – MNAR missing data mechanism

This macro selects records based on the MNAR missing data mechanism with 3 variants: MNAR Left, MNAR Middle, and MNAR Right. These 3 variants correspond to a missing data mechanism whereby SOFA scores in the left, middle, and right sides of the empirical SOFA distribution respectively are selected for deletion.

```

/*   Quantiles
                                LEFT  MID  RIGHT
                                Rate  Rate  Rate
Q1:  0-5      469      52    00    00
Q2:  6-8      536      48    50    00
Q3:  9-11     428      00    50    48
Q4: 12-24     497      00    00    52

Type= can be LEFT, MID, or RIGHT
*/

%macro mnar(in=,out=,type=,percent=);
  data _NULL_; *** output the # obs to variable nrows;
  if 0 then set &in nobs=n;
  call symputx('nrows',n);
  stop;
run;
%let n_SetMissing = %sysevalf(&nrows *
                             (&percent/100),ceil);

%if &type=LEFT %then %do;
  %let Q1rate = 0.52;
  %let Q2rate = 0.48;
  %let n_Q1 = %sysevalf(&Q1rate * (&percent/100) *
                       &nrows, ceil);
  %let n_Q2 = %sysevalf(&Q2rate * (&percent/100) *
                       &nrows, ceil);

  proc surveyselect data=&in(where=(0<=SOFA<=5))
    out=mnar1 method=SRS outall
    sampsize=&n_Q1
    noprint;
  run;
  proc surveyselect data=&in(where=(6<=SOFA<=8))
    out=mnar2 method=SRS outall
    sampsize=&n_Q2
    noprint;
  run;

```



```

    data mnar3;
      set &in;
      where SOFA>=9;
      selected=0;
    run;
  %end;

  %if &type=MID %then %do;
    %let Q2rate = 0.50;
    %let Q3rate = 0.50;
    %let n_Q2 = %sysevalf(&Q2rate * (&percent/100) *
                          &nrows, ceil);
    %let n_Q3 = %sysevalf(&Q3rate * (&percent/100) *
                          &nrows, ceil);

    proc surveysselect data=&in(where=(6<=SOFA<=8))
      out=mnar1 method=SRS outall
      sampsize=&n_Q2
      noprint;
    run;
    proc surveysselect data=&in(where=(9<=SOFA<=11))
      out=mnar2 method=SRS outall
      sampsize=&n_Q3
      noprint;
    run;
    data mnar3;
      set &in;
      where (SOFA <= 5) or (SOFA >= 12);
      selected=0;
    run;
  %end;

  %if &type=RIGHT %then %do;
    %let Q3rate = 0.48;
    %let Q4rate = 0.52;
    %let n_Q3 = %sysevalf(&Q3rate * (&percent/100) *
                          &nrows, ceil);
    %let n_Q4 = %sysevalf(&Q4rate * (&percent/100) *
                          &nrows, ceil);

    proc surveysselect data=&in(where=(9<=SOFA<=11))
      out=mnar1 method=SRS outall
      sampsize=&n_Q3
      noprint;
    run;
  %end;

```

```
proc surveystest data=&in(where=(12<=SOFA<=24))
  out=mnar2 method=SRS outall
  sampsize=&n_Q4
  noprint;
run;
data mnar3;
  set &in;
  where SOFA <= 8;
  selected=0;
run;
%end;

data &out;
  set mnar1 mnar2 mnar3;
run;

proc delete data=mnar1; run;
proc delete data=mnar2; run;
proc delete data=mnar3; run;

%mend;
```

Multiple Imputation – Composite (Aim 1)

```
proc mi data=&SimTable nimpute=25 out=temp_Method4;
  by sim_run;
  class Age_Group Male Race2 payor_group2;
    * Died should not be a class variable, per prior
    literature;
  var SOFA
    Age_Group Male Race2 payor_group2
    Died ICU_LOS TotalCharges /* Outcomes */
    SOFA_CNS SOFA_Coag SOFA_Hep SOFA_Ren SOFA_Resp
    CharlsScore;
  fcs reg(SOFA);
  transform log(TotalCharges) log(ICU_LOS);
  ods output ModelInfo=MI_Seeds1
    VarianceInfo=MI_Variance1
    ParameterEstimates=MI_Parms1;
run;

proc sql;
  create index sim_mi on temp_Method4 (sim_run,
    _imputation_ );
quit;
```

Multiple Imputation – Component (Aim 2)

```

proc mi data=&SimTable nimpute=25 out=temp_Method4;
  by sim_run;
  class Age_Group Male Race2 payor_group2;
  var Age_Group Male Race2 payor_group2
      Died ICU_LOS TotalCharges          /* Outcomes */
      SOFA_CNS SOFA_Card SOFA_Coag SOFA_Hep SOFA_Ren
      SOFA_Resp CharlsScore;
  fcs reg(SOFA_CNS SOFA_Card SOFA_Coag SOFA_Hep SOFA_Ren
          SOFA_Resp);
  transform log(TotalCharges) log(ICU_LOS);
  ods output ModelInfo=MI_Seeds1
             VarianceInfo=MI_Variancel
             ParameterEstimates=MI_Parms1;
run;

data temp_Method4;          *Set MIN component score to zero ;
set temp_Method4;
if SOFA_CNS < 0 then SOFA_CNS = 0;
if SOFA_Card < 0 then SOFA_Card = 0;
if SOFA_Coag < 0 then SOFA_Coag = 0;
if SOFA_Hep < 0 then SOFA_Hep = 0;
if SOFA_Ren < 0 then SOFA_Ren = 0;
if SOFA_Resp < 0 then SOFA_Resp = 0;
SOFA = SOFA_CNS + SOFA_Card + SOFA_Coag + SOFA_Hep +
      SOFA_Ren + SOFA_Resp;
run;

```

REFERENCES

1. Halpern NA, Pastores SM, Greenstein RJ. (2004). Critical care medicine in the United States 1985-2000: an analysis of bed numbers, use, and costs. *Crit Care Med*, 32(6):1254-1259. PubMed PMID: 15187502.
2. Halpern NA, Pastores SM. (2010). Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med*, 38(1):65-71. doi: 10.1097/CCM.0b013e3181b090d0. PubMed PMID: 19730257.
3. Carson SS, Cox CE, Holmes GM, Howard A, Carey TS. (2006). The changing epidemiology of mechanical ventilation: a population-based study. *J Intensive Care Med*, 21(3):173-182. doi: 10.1177/0885066605282784. PubMed PMID: 16672639.
4. Jones C, Griffiths RD, Humphris G, Skirrow PM. (2001). Memory, delusions, and the development of acute posttraumatic stress disorder-related symptoms after intensive care. *Crit Care Med*, 29(3):573-580. PubMed PMID: 11373423.
5. Marra A, Pandharipande PP, Patel MB. (2017). Intensive Care Unit Delirium and Intensive Care Unit-Related Posttraumatic Stress Disorder. *Surgical Clinics of North America*, 97(6):1215-1235. doi: 10.1016/j.suc.2017.07.008. PubMed PMID: 29132506.
6. Bienvenu OJ, Gerstenblith TA. (2017). Posttraumatic Stress Disorder Phenomena After Critical Illness. *Critical Care Clinics*, 33(3):649-658. doi: 10.1016/j.ccc.2017.03.006. PubMed PMID: 28601139.
7. Treggiari MM, Romand J-A, Yanez ND, Deem SA, Goldberg J, Hudson L, Heidegger C-P, Weiss NS. (2009). Randomized trial of light versus deep sedation on mental health after critical illness*. *Critical Care Medicine*, 37(9):2527-2534. doi: 10.1097/CCM.0b013e3181a5689f. PubMed PMID: 00003246-200909000-00005.
8. Marra A, Ely EW, Pandharipande PP, Patel MB. (2017). The ABCDEF Bundle in Critical Care. *Critical Care Clinics*, 33(2):225-243. doi: 10.1016/j.ccc.2016.12.005. PubMed PMID: 28284292.
9. McGiffin JN, Galatzer-Levy IR, Bonanno GA. (2016). Is the intensive care unit traumatic? What we know and don't know about the intensive care unit and posttraumatic stress responses. *Rehabilitation Psychology*, 61(2):120-131. doi: 10.1037/rep0000073. PubMed PMID: 27196855.

10. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ : British Medical Journal*, 350. doi: 10.1136/bmj.h391.
11. Strand K, Flaatten H. (2008). Severity scoring in the ICU: a review. *Acta Anaesthesiologica Scandinavica*, 52(4):467-478. doi: 10.1111/j.1399-6576.2008.01586.x.
12. Schafer JL. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall. 444 p.
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gotsche PC, Vandenbroucke JP. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*, 147(8):573-577. PubMed PMID: 17938396.
14. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, Sorensen HT, von Elm E, Langan SM. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*, 12(10):e1001885. doi: 10.1371/journal.pmed.1001885. PubMed PMID: 26440803.
15. International Committee of Medical Journal Editors. (2016). Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals. Available from: <http://www.icmje.org/>.
16. Moher B, Brooks J, Clark MA, Crown WH, Davey P, Hutchins D, Martin BC, Stang P. (2003). A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. *Value Health*, 6(2):90-97. doi: 10.1046/j.1524-4733.2003.00242.x. PubMed PMID: 12641858.
17. Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goettsch W, Madigan D, Makady A, Schneeweiss S, Tarricone R, Wang SV, Watkins J, Mullins CD. (2017). Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value in Health*, 20(8):1003-1008. doi: 10.1016/j.jval.2017.08.3019.
18. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, Gagne JJ, Gini R, Klungel O, Mullins CD, Nguyen MD, Rassen JA, Smeeth L, Sturkenboom M. (2017).

Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. *Value in Health*, 20(8):1009-1022. doi: 10.1016/j.jval.2017.08.3018.

19. PCORI. (2017). *Methodology Standards*. Available from: <https://www.pcori.org/sites/default/files/PCORI-Methodology-Standards.pdf>.
20. Rezvan PH, Lee KJ, Simpson JA. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol*, 15:1-14. doi: 10.1186/s12874-015-0022-1. PubMed PMID: 25880850.
21. Truven Health Analytics. (2016). *Truven Health MarketScan® Research Databases: Commercial Claims and Encounters User Guide*. Ann Arbor, MI.
22. Birt JA, Tan Y, Mozaffarian N. (2017). Sjogren's syndrome: managed care data from a large United States population highlight real-world health care burden and lack of treatment options. *Clin Exp Rheumatol*, 35(1):98-107. PubMed PMID: 27749234.
23. Evans R, Loeb A, Kaye KS, Cher ML, Martin ET. (2017). Infection-Related Hospital Admissions After Prostate Biopsy in United States Men. *Open Forum Infect Dis*, 4(1):1-3. doi: 10.1093/ofid/ofw265. PubMed PMID: 28480258.
24. Lip GY, Hunter TD, Quiroz ME, Ziegler PD, Turakhia MP. (2017). Atrial Fibrillation Diagnosis Timing, Ambulatory ECG Monitoring Utilization, and Risk of Recurrent Stroke. *Circ Cardiovasc Qual Outcomes*, 10(1):1-8. doi: 10.1161/circoutcomes.116.002864. PubMed PMID: 28096204.
25. Nichols CI, Vose JG. (2017). Incidence of Bleeding-Related Complications During Primary Implantation and Replacement of Cardiac Implantable Electronic Devices. *J Am Heart Assoc*, 6(1):1-9. doi: 10.1161/jaha.116.004263. PubMed PMID: 28111362.
26. Qin X, Tangka FK, Guy GP, Jr., Howard DH. (2017). Mammography rates after the 2009 revision to the United States Preventive Services Task Force breast cancer screening recommendation. *Cancer Causes Control*, 28(1):41-48. doi: 10.1007/s10552-016-0835-1. PubMed PMID: 28025762.
27. Saeed MJ, Olsen MA, Powderly WG, Presti RM. (2017). Diabetes Mellitus is Associated With Higher Risk of Developing Decompensated Cirrhosis in Chronic Hepatitis C Patients. *J Clin Gastroenterol*, 51(1):70-76. doi: 10.1097/mcg.0000000000000566. PubMed PMID: 27306942.

28. Sajisevi M, Schulz K, Cyr DD, Wojdyla D, Rosenfeld RM, Tucci D, Witsell DL. (2017). Nonadherence to Guideline Recommendations for Tympanostomy Tube Insertion in Children Based on Mega-database Claims Analysis. *Otolaryngol Head Neck Surg*, 156(1):87-95. doi: 10.1177/0194599816669499. PubMed PMID: 27625028.
29. Solid CA, Peter SA, Natwick T, Guo H, Collins AJ, Arduino JM. (2017). Impact of Renal Disease on Patients with Hepatitis C: A Retrospective Analysis of Disease Burden, Clinical Outcomes, and Health Care Utilization and Cost. *Nephron*, 136(2):54-61. doi: 10.1159/000454684. PubMed PMID: 28214902.
30. Stephens JR, Steiner MJ, DeJong N, Rodean J, Hall M, Richardson T, Berry JG. (2017). Healthcare Utilization and Spending for Constipation in Children With Versus Without Complex Chronic Conditions. *J Pediatr Gastroenterol Nutr*, 64(1):31-36. doi: 10.1097/mpg.0000000000001210. PubMed PMID: 27070656.
31. Tao G, Patel C, Hoover KW. (2017). Updated Estimates of Ectopic Pregnancy among Commercially and Medicaid-Insured Women in the United States, 2002-2013. *South Med J*, 110(1):18-24. doi: 10.14423/smj.0000000000000594. PubMed PMID: 28052169.
32. Wallace L, Kadakia A. (2017). Buprenorphine transdermal system utilization. *Postgrad Med*, 129(1):81-86. doi: 10.1080/00325481.2017.1267537. PubMed PMID: 27901359.
33. Wu JJ, Guerin A, Sundaram M, Dea K, Cloutier M, Mulani P. (2017). Cardiovascular event risk assessment in psoriasis patients treated with tumor necrosis factor-alpha inhibitors versus methotrexate. *J Am Acad Dermatol*, 76(1):81-90. doi: 10.1016/j.jaad.2016.07.042. PubMed PMID: 27894789.
34. Zhang D, Johnson K, Newransky C, Acosta CJ. (2017). Herpes Zoster Vaccine Coverage in Older Adults in the U.S., 2007-2013. *Am J Prev Med*, 52(1):e17-e23. doi: 10.1016/j.amepre.2016.08.029. PubMed PMID: 28340974.
35. Brittan M, Richardson T, Kenyon C, Sills MR, Fieldston E, Hall M, Fox D, Shah S, Berry J. (2017). Association between Postdischarge Oral Corticosteroid Prescription Fills and Readmission in Children with Asthma. *J Pediatr*, 180:163-169.e161. doi: 10.1016/j.jpeds.2016.09.034. PubMed PMID: 27769549.
36. Herzog MM, Marshall SW, Lund JL, Pate V, Spang JT. (2017). Cost of Outpatient Arthroscopic Anterior Cruciate Ligament Reconstruction Among Commercially Insured

- Patients in the United States, 2005-2013. *Orthop J Sports Med*, 5(1):1-8. doi: 10.1177/2325967116684776. PubMed PMID: 28210655.
37. Millman AJ, Reynolds S, Duffy J, Chen J, Gargiullo P, Fry AM. (2017). Hospitalizations within 14 days of vaccination among pediatric recipients of the live attenuated influenza vaccine, United States 2010-2012. *Vaccine*, 35(4):529-535. doi: 10.1016/j.vaccine.2016.12.033. PubMed PMID: 28041779.
38. Ullal AJ, Kaiser DW, Fan J, Schmitt SK, Than CT, Winkelmayr WC, Heidenreich PA, Piccini JP, Perez MV, Wang PJ, Turakhia MP. (2017). Safety and Clinical Outcomes of Catheter Ablation of Atrial Fibrillation in Patients With Chronic Kidney Disease. *J Cardiovasc Electrophysiol*, 28(1):39-48. doi: 10.1111/jce.13118. PubMed PMID: 27782345.
39. Wu H, Mendoza MC, Huang YA, Hayes T, Smith DK, Hoover KW. (2017). Uptake of HIV Preexposure Prophylaxis Among Commercially Insured Persons-United States, 2010-2014. *Clin Infect Dis*, 64(2):144-149. doi: 10.1093/cid/ciw701. PubMed PMID: 27986691.
40. Rosenbaum PR, Rubin DB. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41-55. doi: 10.1093/biomet/70.1.41.
41. Rubin DB. (1997). Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*, 127(8 Pt 2):757-763. PubMed PMID: 9382394.
42. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs*, 24(5):1103-1117.
43. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, 62(8):1120-1127.
44. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. (2009). Use of electronic medical records for health outcomes research: A literature review. *Medical Care Research and Review*, 66(6):611-638.
45. eMERGE. (2014). *About eMERGE*. [Accessed 10/31/2017]. Available from: <https://emerge.mc.vanderbilt.edu/about-emerge/>.

46. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ. (2011). Electronic medical records for genetic research: results of the eMERGE consortium. *Science Translational Medicine*, 3(79):79re71-79re71.
47. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB, Lingren T, Thompson WK, Savova G, Haines J, Roden DM, Harris PA, Denny JC. (2016). PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc*, 23(6):1046-1052. doi: 10.1093/jamia/ocv202. PubMed PMID: 27026615.
48. Shortliffe EH. (1999). The evolution of electronic medical records. *Academic Medicine*, 74(4):414-419.
49. Grippi MA. (2015). *Respiratory Failure: An Overview*. In: Grippi MA, Elias JA, Fishman JA, Kotloff RM, Pack AI, Senior RM, Siegel MD, editors. *Fishman's Pulmonary Diseases and Disorders*, 5e. New York, NY: McGraw-Hill Education. p.
50. Matthay MA, Slutsky AS. (2016). *Acute Respiratory Failure*. In: Goldman-Cecil Medicine [Internet]. Philadelphia, PA: Elsevier-Saunders. 25th. [655-664].
51. Halle MJ, Levant S, DeFrances CJ. (2013). *Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010*. Hyattsville, MD: National Center for Health Statistics.
52. Kochanek K, Murphy S, Xu J, Tejada-Vera B. (2016). *Deaths: Final data for 2014. National vital statistics reports*. Hyattsville, MD: National Center for Health Statistics, June 30, 2016. Report No.: Contract No.: 4.
53. Hayman WR, Leuthner SR, Laventhal NT, Brousseau DC, Lagatta JM. (2015). Cost comparison of mechanically ventilated patients across the age span. *Journal of Perinatology*, 35(12):1020-1026. doi: 10.1038/jp.2015.131. PubMed PMID: 26468935.
54. Khatutsky G, Ormond C, Wiener J, Greene A, Johnson R, Jessup E, Vreeland E, Sengupta M, Caffrey C, Harris-Kojetin L. (2016). *Residential care communities and their residents in 2010: A national portrait*. In: Department of Health and Human Services, editor. Hyattsville, MD: National Center for Health Statistics.

55. Farias JA, Fernandez A, Monteverde E, Flores JC, Baltodano A, Menchaca A, Poterala R, Panico F, Johnson M, von Dessauer B, Donoso A, Zavala I, Zavala C, Troster E, Pena Y, Flamenco C, Almeida H, Nilda V, Esteban A. (2012). Mechanical ventilation in pediatric intensive care units during the season for acute lower respiratory infection: a multicenter study. *Pediatric Critical Care Medicine*, 13(2):158-164. doi: 10.1097/PCC.0b013e3182257b82. PubMed PMID: 21725275.
56. Aitken LM, Bucknall T, Kent B, Mitchell M, Burmeister E, Keogh SJ. (2015). Protocol-directed sedation versus non-protocol-directed sedation to reduce duration of mechanical ventilation in mechanically ventilated intensive care patients. *Cochrane Database Syst Rev*, 1:Cd009771. doi: 10.1002/14651858.CD009771.pub2. PubMed PMID: 25562750.
57. Esteban A, Anzueto A, Frutos F, Alia I, Brochard L, Stewart TE, Benito S, Epstein SK, Apezteguia C, Nightingale P, Arroliga AC, Tobin MJ. (2002). Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study. *Jama*, 287(3):345-355. PubMed PMID: 11790214.
58. Metnitz PG, Metnitz B, Moreno RP, Bauer P, Del Sorbo L, Hoermann C, de Carvalho SA, Ranieri VM. (2009). Epidemiology of mechanical ventilation: analysis of the SAPS 3 database. *Intensive Care Med*, 35(5):816-825. doi: 10.1007/s00134-009-1449-9. PubMed PMID: 19288079.
59. Charlson ME, Pompei P, Ales KL, MacKenzie CR. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*, 40(5):373-383. PubMed PMID: 3558716.
60. Deyo RA, Cherkin DC, Ciol MA. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*, 45(6):613-619. PubMed PMID: 1607900.
61. Elixhauser A, Steiner C, Harris DR, Coffey RM. (1998). Comorbidity measures for use with administrative data. *Med Care*, 36(1):8-27. PubMed PMID: 9431328.
62. Ladha KS, Zhao K, Quraishi SA, Kurth T, Eikermann M, Kaafarani HM, Klein EN, Seethala R, Lee J. (2015). The Deyo-Charlson and Elixhauser-van Walraven Comorbidity Indices as predictors of mortality in critically ill patients. *BMJ Open*, 5(9):e008990. doi: 10.1136/bmjopen-2015-008990. PubMed PMID: 26351192.
63. Stavem K, Hoel H, Skjaker SA, Haagensen R. (2017). Charlson comorbidity index derived from chart review or administrative data: agreement and prediction of mortality

- in intensive care patients. *Clinical Epidemiology*, 9:311-320. doi: 10.2147/clep.s133624. PubMed PMID: 28652813.
64. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, Reinhart CK, Suter PM, Thijs LG. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med*, 22(7):707-710. PubMed PMID: 8844239.
65. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Cooper-Smith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent JL, Angus DC. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama*, 315(8):801-810. doi: 10.1001/jama.2016.0287. PubMed PMID: 26903338.
66. Office of the Assistant Secretary for Preparedness & Response. (2017). *SOFA Score: What it is and how to use it in triage*. In: U.S. Department of Health & Human Services, editor.
67. Knox DB, Lanspa MJ, Pratt CM, Kuttler KG, Jones JP, Brown SM. (2014). Glasgow Coma Scale score dominates the association between admission Sequential Organ Failure Assessment score and 30-day mortality in a mixed intensive care unit population. *J Crit Care*, 29(5):780-785. doi: 10.1016/j.jcrc.2014.05.009. PubMed PMID: 25012961.
68. Stoller JK, Hill NS. (2016). *Respiratory Monitoring in Critical Care*. In: Goldman-Cecil Medicine [Internet]. Philadelphia, PA: Elsevier-Saunders. 25th. [652-655].
69. Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, Slutsky AS. (2012). Acute respiratory distress syndrome: the Berlin Definition. *Jama*, 307(23):2526-2533. doi: 10.1001/jama.2012.5669. PubMed PMID: 22797452.
70. Abrams CS. (2016). *Thrombocytopenia*. In: Goldman-Cecil Medicine [Internet]. Philadelphia, PA: Elsevier-Saunders. 25th. [1159-1167].
71. Berk PD, Korenblat KM. (2016). *Approach to the Patient with Jaundice or Abnormal Liver Tests*. In: Goldman-Cecil Medicine [Internet]. Philadelphia, PA: Elsevier-Saunders. 25th. [983-993].

72. Wan Z, Wu Y, Yi J, You S, Liu H, Sun Z, Zhu B, Zang H, Li C, Liu F, Li D, Mao Y, Xin S. (2015). Combining serum cystatin C with total bilirubin improves short-term mortality prediction in patients with HBV-related acute-on-chronic liver failure. *PLoS One*, *10*(1):e0116968. doi: 10.1371/journal.pone.0116968. PubMed PMID: 25629773.
73. Ong KL, Allison MA, Cheung BM, Wu BJ, Barter PJ, Rye KA. (2014). The relationship between total bilirubin levels and total mortality in older adults: the United States National Health and Nutrition Examination Survey (NHANES) 1999-2004. *PLoS One*, *9*(4):e94479. doi: 10.1371/journal.pone.0094479. PubMed PMID: 24728477.
74. Zheng MH, Shi KQ, Lin XF, Xiao DD, Chen LL, Liu WY, Fan YC, Chen YP. (2013). A model to predict 3-month mortality risk of acute-on-chronic hepatitis B liver failure using artificial neural network. *J Viral Hepat*, *20*(4):248-255. doi: 10.1111/j.1365-2893.2012.01647.x. PubMed PMID: 23490369.
75. Su HH, Kao CM, Lin YC, Lin YC, Kao CC, Chen HH, Hsu CC, Chen KC, Peng CC, Wu MS. (2017). Relationship between serum total bilirubin levels and mortality in uremia patients undergoing long-term hemodialysis: A nationwide cohort study. *Atherosclerosis*, *265*:155-161. doi: 10.1016/j.atherosclerosis.2017.09.001. PubMed PMID: 28892712.
76. Rivers EP. (2016). *Approach to the Patient with Shock*. In: Goldman-Cecil Medicine [Internet]. Philadelphia, PA: Elsevier-Saunders. 25th. [672-681].
77. Teasdale G, Jennett B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet*, *2*(7872):81-84. PubMed PMID: 4136544.
78. Teasdale G, Maas A, Lecky F, Manley G, Stocchetti N, Murray G. The Glasgow Coma Scale at 40 years: standing the test of time. *The Lancet Neurology*, *13*(8):844-854. doi: 10.1016/S1474-4422(14)70120-6.
79. Centers for Medicare & Medicaid Services, National Center for Health Statistics. (2017). *ICD-10-CM Official Guidelines for Coding and Reporting, FY 2018*. Available from: https://www.cdc.gov/nchs/data/icd/10cmguidelines2011_FINAL.pdf.
80. Molitoris BA. (2016). *Acute Kidney Injury*. In: Goldman-Cecil Medicine [Internet]. Philadelphia, PA: Elsevier-Saunders. 25th. [778-783].
81. Susantitaphong P, Cruz DN, Cerda J, Abulfaraj M, Alqahtani F, Koulouridis I, Jaber BL. (2013). World incidence of AKI: a meta-analysis. *Clinical Journal of the American Society of Nephrology*, *8*(9):1482-1493.

82. Hoste EA, Clermont G, Kersten A, Venkataraman R, Angus DC, De Bacquer D, Kellum JA. (2006). RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis. *Crit Care*, 10(3):R73. doi: 10.1186/cc4915. PubMed PMID: 16696865.
83. Barrantes F, Tian J, Vazquez R, Amoateng-Adjepong Y, Manthous CA. (2008). Acute kidney injury criteria predict outcomes of critically ill patients. *Crit Care Med*, 36(5):1397-1403. doi: 10.1097/CCM.0b013e318168fbe0. PubMed PMID: 18434915.
84. Vincent JL, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, Sprung CL, Colardyn F, Blecher S. (1998). Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. *Crit Care Med*, 26(11):1793-1800. PubMed PMID: 9824069.
85. Matics TJ, Sanchez-Pinto LN. (2017). Adaptation and Validation of a Pediatric Sequential Organ Failure Assessment Score and Evaluation of the Sepsis-3 Definitions in Critically Ill Children. *JAMA Pediatr*:e172352. doi: 10.1001/jamapediatrics.2017.2352. PubMed PMID: 28783810.
86. Forte DN, Ranzani OT, Stape N, Taniguchi LU, Toledo-Maciel A, Park M. (2007). APACHE II and SOFA scores for intensive care and hospital outcome prediction in oncologic patients. *Critical Care*, 11(Suppl 3):P93-P93. doi: 10.1186/cc5880. PubMed PMID: PMC3301220.
87. Elsayed FG, Sholkamy AA, Elshazli M, Elshafie M, Naguib M. (2015). Comparison of different scoring systems in predicting short-term mortality after liver transplantation. *Transplantation Proceedings*, 47(4):1207-1210. doi: 10.1016/j.transproceed.2014.11.067. PubMed PMID: 26036555.
88. Hwang SY, Lee JH, Lee YH, Hong CK, Sung AJ, Choi YC. (2012). Comparison of the Sequential Organ Failure Assessment, Acute Physiology and Chronic Health Evaluation II scoring system, and Trauma and Injury Severity Score method for predicting the outcomes of intensive care unit trauma patients. *Am J Emerg Med*, 30(5):749-753. doi: 10.1016/j.ajem.2011.05.022. PubMed PMID: 21802884.
89. Demandt AMP, Geerse DA, Janssen BJP, Winkens B, Schouten HC, van Mook W. (2017). The prognostic value of a trend in modified SOFA score for patients with hematological malignancies in the intensive care unit. *European Journal of Haematology*, 99(4):315-322. doi: 10.1111/ejh.12919. PubMed PMID: 28656589.

90. Mazzola P, Bellelli G, Perego S, Zambon A, Mazzone A, Bruni AA, Annoni G. (2013). The sequential organ failure assessment score predicts 30-day mortality in a geriatric acute care setting. *J Gerontol A Biol Sci Med Sci*, 68(10):1291-1295. doi: 10.1093/gerona/glt020. PubMed PMID: 23580741.
91. Kantanen AM, Kalviainen R, Parviainen I, Ala-Peijari M, Backlund T, Koskenkari J, Laitio R, Reinikainen M. (2017). Predictors of hospital and one-year mortality in intensive care patients with refractory status epilepticus: a population-based study. *Crit Care*, 21(1):71. doi: 10.1186/s13054-017-1661-x. PubMed PMID: 28330483.
92. Oeyen S, Vermeulen K, Benoit D, Annemans L, Decruyenaere J. (2017). Development of a prediction model for long-term quality of life in critically ill patients. *J Crit Care*, 43:133-138. doi: 10.1016/j.jcrc.2017.09.006. PubMed PMID: 28892669.
93. Neto AS, Barbas CSV, Simonis FD, Artigas-Raventos A, Canet J, Determann RM, Anstey J, Hedenstierna G, Hemmes SNT, Hermans G, Hiesmayr M, Hollmann MW, Jaber S, Martin-Loeches I, Mills GH, Pearse RM, Putensen C, Schmid W, Severgnini P, Smith R, Treschan TA, Tschernko EM, Melo MFV, Wrigge H, de Abreu MG, Pelosi P, Schultz MJ. (2016). Epidemiological characteristics, practice of ventilation, and clinical outcome in patients at risk of acute respiratory distress syndrome in intensive care units from 16 countries (PRoVENT): an international, multicentre, prospective study. *Lancet Respir Med*, 4(11):882-893. doi: 10.1016/s2213-2600(16)30305-8. PubMed PMID: 27717861.
94. Ferreira FL, Bota DP, Bross A, Melot C, Vincent JL. (2001). Serial evaluation of the SOFA score to predict outcome in critically ill patients. *Jama*, 286(14):1754-1758. PubMed PMID: 11594901.
95. Rubin DB. (1976). Inference and missing data. *Biometrika*, 63(3):581-592.
96. Little RJA, Rubin DB. (2002). *Statistical analysis with missing data*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons.
97. Allison PD. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications. 91 p.
98. Rubin DB. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons. 258 p.

99. Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. (2015). Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. *International Journal of Epidemiology*, 44(3):937-945. doi: 10.1093/ije/dyv035.
100. Ono M, Miller HP. (1969). *Income nonresponses in the current population survey*: US Bureau of the Census.
101. Graham JW. (2009). Missing data analysis: making it work in the real world. *Annu Rev Psychol*, 60:549-576. doi: 10.1146/annurev.psych.58.110405.085530. PubMed PMID: 18652544.
102. Bell ML, Fairclough DL, Fiero MH, Butow PN. (2016). Handling missing items in the Hospital Anxiety and Depression Scale (HADS): a simulation study. *BMC Res Notes*, 9(1):479. doi: 10.1186/s13104-016-2284-z. PubMed PMID: 27770833.
103. Little RJA. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83(404):1198-1202. doi: 10.1080/01621459.1988.10478722.
104. Allison PD. (2009). *Missing Data*. In: Milsap RE, Maydeu-Olivares A, editors. *The SAGE Handbook of Quantitative Methods in Psychology*. London: SAGE Publications Ltd. p.
105. Park T, Davis CS. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49(2):631-638. PubMed PMID: 8369395.
106. Park T, Lee SY. (1997). A test of missing completely at random for longitudinal data with missing observations. *Stat Med*, 16(16):1859-1871. PubMed PMID: 9280038.
107. Heitjan DF, Basu S. (1996). Distinguishing missing at random and missing completely at random. *American Statistician*, 50(3):207.
108. Yeatts SD, Martin RH. (2015). What is missing from my missing data plan? *Stroke*, 46(6):e130-132. doi: 10.1161/strokeaha.115.007984. PubMed PMID: 25953373.
109. Knol MJ, Janssen KJ, Donders AR, Egberts AC, Heerdink ER, Grobbee DE, Moons KG, Geerlings MI. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin*

- Epidemiol*, 63(7):728-736. doi: 10.1016/j.jclinepi.2009.08.028. PubMed PMID: 20346625.
110. Cohen J, Cohen P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum.
 111. Jones MP. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*, 91(433):222-230. doi: 10.2307/2291399.
 112. Lagu T, Lindenauer PK, Rothberg MB, Nathanson BH, Pekow PS, Steingrub JS, Higgins TL. (2011). Development and validation of a model that uses enhanced administrative data to predict mortality in patients with sepsis. *Crit Care Med*, 39(11):2425-2430. doi: 10.1097/CCM.0b013e31822572e3. PubMed PMID: 22005222.
 113. Rubin DB. (1996). Multiple Imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473-489.
 114. Eekhout I, de Vet HC, Twisk JW, Brand JP, de Boer MR, Heymans MW. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *J Clin Epidemiol*, 67(3):335-342. doi: 10.1016/j.jclinepi.2013.09.009. PubMed PMID: 24291505.
 115. Allison PD. (2002). *Missing data: Quantitative applications in the social sciences*: British Psychological Society. 193 p.
 116. Little RJA. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3):287-296. doi: 10.1080/07350015.1988.10509663.
 117. Rubin DB. (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 4(1):87-94. doi: 10.1080/07350015.1986.10509497.
 118. U.S. Census Bureau. (2017). *Center for Statistical Research & Methodology (CSRM): Missing Data, Edit, and Imputation*. [Accessed 10/1/2017]. Available from: <https://www.census.gov/srd/csr/missingData.html>.

119. Andridge RR, Little RJA. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International statistical review*, 78(1):40-64. doi: 10.1111/j.1751-5823.2010.00103.x. PubMed PMID: PMC3130338.
120. Bounthavong M, Watanabe JH, Sullivan KM. (2015). Approach to addressing missing data for electronic medical records and pharmacy claims data research. *Pharmacotherapy: The Journal of Human Pharmacology & Drug Therapy*, 35(4):380-387. doi: <https://dx.doi.org/10.1002/phar.1569>. PubMed PMID: 25884526.
121. van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*, 16(3):219-242. doi: 10.1177/0962280206074463. PubMed PMID: 17621469.
122. Heitjan DF, Little RJ. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*:13-29.
123. Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE, Moons KG. (2010). Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol*, 63(7):721-727. doi: 10.1016/j.jclinepi.2009.12.008. PubMed PMID: 20338724.
124. Welch CA, Petersen I, Bartlett JW, White IR, Marston L, Morris RW, Nazareth I, Walters K, Carpenter J. (2014). Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, 33(21):3725-3737. doi: <https://dx.doi.org/10.1002/sim.6184>. PubMed PMID: 24782349.
125. Lee KJ, Carlin JB. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5):624-632. doi: 10.1093/aje/kwp425. PubMed PMID: 20106935.
126. Fitzmaurice GM, Kenward MG, Molenberghs G, Verbeke G, Tsiatis AA. (2015). *Missing data: Introduction and statistical preliminaries*. In: Molenberghs G, Fitzmaurice GM, Kenward MG, Tsiatis AA, Verbeke G, editors. *Handbook of missing data methodology*. New York, NY: CRC Press. p. 3-22.
127. Dempster AP, Laird NM, Rubin DB. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*:1-38.

128. Binquet C, VERRET C, CHENE G, SALMI L, LETENNEUR L, PALMER G, HAJJAR M, SALAMON R. (1998). Principaux logiciels statistiques utilisables en épidémiologie. *Revue d'épidémiologie et de santé publique*, 46(4):329-336.
129. Yergens DW, Dutton DJ, Patten SB. (2014). An overview of the statistical methods reported by studies using the Canadian community health survey. *BMC Medical Research Methodology*, 14:1-7. doi: 10.1186/1471-2288-14-15. PubMed PMID: PMC3922729.
130. Dembe AE, Partridge JS, Geist LC. (2011). Statistical software applications used in health services research: analysis of published studies in the U.S. *BMC Health Services Research*, 11:1-6. doi: 10.1186/1472-6963-11-252. PubMed PMID: PMC3205033.
131. Research Triangle Institute. (2017). *About SUDAAN*. [Accessed 9/13/2017]. Available from: http://sudaansupport.rti.org/page.cfm/About_SUDAAN.
132. PCORI Data Quality and Missing Data Workgroup. (2015). *Data Quality and Missing Data in Patient-Centered Outcomes Research Using EMR/Claims Data Meeting Summary*. [Accessed 8/31/2017]. Available from: www.pcori.org/sites/default/files/PCORI-Data-Quality-and-Missing-Data-Workgroup-Summary-121015.pdf.
133. Medical University of South Carolina. (n.d.). *The Enterprise Data Warehouse*. [Accessed 10/20/2017]. Available from: <http://academicdepartments.musc.edu/edw/whytheEDW>.
134. Medical University of South Carolina. (n.d.). *Accessing the EDW: Research*. [Accessed 10/20/2017]. Available from: <http://academicdepartments.musc.edu/edw/accessingtheEDW/Research/>.
135. Medical University of South Carolina. (2017). *MUSC Data Warehouse Status Run at: Fri Oct 20 12:21:05 2017*. [Accessed 10/20/2017]. Available from: <http://timon.musc.edu/lbg3/curstatus.html>.
136. Medical University of South Carolina. (n.d.). *Biomedical Informatics Center: Research Data Overview*. [Accessed 10/20/2017]. Available from: <http://academicdepartments.musc.edu/bmic/ResearchDataOverview.html>.

137. Burton A, Altman DG, Royston P, Holder RL. (2006). The design of simulation studies in medical statistics. *Stat Med*, 25(24):4279-4292. doi: 10.1002/sim.2673. PubMed PMID: 16947139.
138. Schulz KF, Grimes DA. (2002). Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet*, 359(9308):781-785. doi: 10.1016/s0140-6736(02)07882-0. PubMed PMID: 11888606.
139. Bennett DA. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5):464-469. PubMed PMID: 11688629.
140. Molenberghs G, Beunckens C, Sotto C, Kenward MG. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371-388.
141. Hosmer DW, Lemeshow S, Sturdivant RX. (2013). *Applied logistic regression*. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc. 500 p.
142. Brand JP, van Buuren S, Groothuis-Oudshoorn K, Gelsema ES. (2003). A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1):36-45.
143. Ambler G, Omar RZ, Royston P. (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research*, 16(3):277-298.
144. Bondarenko I, Raghunathan T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, 35(17):3007-3020. doi: 10.1002/sim.6926.
145. Plumpton CO, Morris T, Hughes DA, White IR. (2016). Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Res Notes*, 9:45. doi: 10.1186/s13104-016-1853-5. PubMed PMID: 26809812.
146. Allison PD. (2012). Paper 312-2012: Handling missing data by maximum likelihood. *SAS Global Forum 2012* [Internet]. Available from: www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.
147. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS, Escobar GJ, Angus DC. (2016).

- Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama*, 315(8):762-774. doi: 10.1001/jama.2016.0288. PubMed PMID: 26903335.
148. Bodner TE. (2008). What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4):651-675. doi: 10.1080/10705510802339072.
149. Herritt B, Chaudhuri D, Thavorn K, Kubelik D, Kyeremanteng K. (2017). Early vs. late tracheostomy in intensive care settings: Impact on ICU and hospital costs. *J Crit Care*, 44:285-288. doi: 10.1016/j.jcrc.2017.11.037. PubMed PMID: 29223743.
150. Yoo BK, Kim M, Sasaki T, Hoch JS, Marcin JP. (2018). Selected Use of Telemedicine in Intensive Care Units Based on Severity of Illness Improves Cost-Effectiveness. *Telemed J E Health*, 24(1):21-36. doi: 10.1089/tmj.2017.0069. PubMed PMID: 28661790.
151. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. (2012). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors. New York, NY: Springer.
152. Manning WG, Basu A, Mullahy J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ*, 24(3):465-488. doi: 10.1016/j.jhealeco.2004.09.011. PubMed PMID: 15811539.
153. Lilly CM, Zuckerman IH, Badawi O, Riker RR. (2011). Benchmark data from more than 240,000 adults that reflect the current practice of critical care in the United States. *Chest*, 140(5):1232-1242. doi: 10.1378/chest.11-0718. PubMed PMID: 21868469.
154. SAS Institute Inc. (2016). *SAS® 9.4 Functions and CALL Routines: Reference*. Fifth ed. Cary, NC: SAS Institute Inc. 1236 p.
155. Matsumoto M, Nishimura T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3-30.
156. Shi L. (2008). *Health services research methods*. Clifton Park, NY: Delmar Cengage Learning, 481 p.

157. van Buuren S. (2010). Item imputation without specifying scale structure. *Methodology*, 6(1):31-36. doi: 10.1027/1614-2241/a000004.
158. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. (2012). Prognostic indices for older adults: a systematic review. *Jama*, 307(2):182-192. doi: 10.1001/jama.2011.1966. PubMed PMID: 22235089.
159. Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035. doi: 10.1038/sdata.2016.35.