ASSESSING THE IMPACT OF MISSING DATA ON HOSPITAL PERFORMANCE PROFILING

By

Michael P. Thompson

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Epidemiology – Doctor of Philosophy

2015

ABSTRACT

ASSESSING THE IMPACT OF MISSING DATA ON HOSPITAL PERFORMANCE PROFILING

By

Michael P. Thompson

Ischemic stroke is a leading cause of mortality, long-term disability, and high healthcare costs in the US. In light of this clinical and financial burden, the Centers for Medicare & Medicaid Services (CMS) has decided to incorporate ischemic stroke measures of 30-day mortality and hospital readmission into its current pay-for-performance program. This decision has come under intense scrutiny, as many clinicians and researchers believe that the current risk adjustment model is inadequate because it does not include a measure of stroke severity. Due to its well-documented importance in individual-level prediction, there is concern that excluding a measure of stroke severity from risk adjustment will lead to incorrect rankings of hospital performance, i.e. hospital profiling.

However, administrative datasets used in CMS currently do not capture a measure of stroke severity, such as the National Institutes of Health Stroke Scale (NIHSS), and in clinical databases which capture NIHSS, it is frequently missing. Little work has been done to asses if the documentation of NIHSS is biased, and if so, what impact bias would have on hospital-level estimates of mortality. In this study, we analyzed data from ischemic stroke patients from an existing stroke registry to identify patterns and characteristics that predict NIHSS documentation at the patient- and hospital-level. Next, we tested for the presence of selection bias in patients with documented NIHSS using the Heckman Selection Model. Finally, using computer simulations, we estimated the impact of missing NIHSS data on hospital profiling of

30-day mortality, under different assumptions about the prevalence and mechanism of missing NIHSS data.

We found that patients with documented NIHSS were, in fact, a biased subsample of all ischemic stroke patients. Documentation of NIHSS was driven by a combination of patient-level and hospital-level factors. At the patient- and hospital-level, analyses suggested that patients with more severe strokes (i.e. increased NIHSS score) were better documented than patients with less severe strokes. These findings were confirmed using the Heckman Selection Model. However, in both analyses, we found that the amount of bias was modest.

In computer simulations, we quantified the impact that missing data would have on the accuracy of hospital ischemic stroke profiling, under different assumptions about how NIHSS data was missing. Any effect of missing NIHSS mechanism was trumped by the impact of missingness on sample size. Because patients with missing NIHSS data were dropped from risk-adjustment models as documentation of NIHSS decreased, the accuracy of hospital risk-standardized mortality rates (RSMRs) estimated by the hierarchical logistic model deteriorated. All of our findings were substantially modified by the hospital ischemic stroke volume, with low volume hospital suffering the worst accuracy. These results are a reflection of the fact that the loss of sample size (either through the documentation rate or hospital volume), increases the amount of shrinkage in RSMR estimates, which makes any random noise more impactful on changes in RSMR. Overall, our findings raise concerns about the addition of NIHSS data into risk adjustment models for hospital-level ischemic stroke outcomes, and illustrate shortcomings in current methodologies used to profile hospitals. It is crucial that data used in risk adjustment for hospital profiling be documented with very high levels of completeness.

This dissertation is dedicated to my family and friends for their support and belief in me throughout my education. Above all, I would like to thank my wife, Megan, for her love and encouragement to pursue my career ambitions.

ACKNOWLEDGEMENTS

I will be forever grateful to my mentor and dissertation committee chair, Dr. Mat Reeves, who pushed me to never settle for good enough and inspired me to pursue greater ambitions. I would also like to thank Dr. Zhehui Luo, Dr. Joseph Gardiner, and Dr. Jim Burke for their commitment and direction as part of my dissertation committee. In addition, I would like to acknowledge the faculty, staff, and fellow students of the Department of Epidemiology and Biostatistics for sharing their knowledge, support, and camaraderie throughout my program.

I would also like to acknowledge Adrienne Nickles of the Michigan Department of Community Health for her assistance in working with the Michigan Stroke Registry, which plays an integral part of my dissertation.

Finally, I am indebted to the Michigan Health & Hospital Association, notably Sam Watson and Steve Levy, for funding my graduate assistantship and giving me the opportunity to learn from and contribute to their continued work in improving the quality of health care across the State of Michigan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AHA/ASA	American Heart Association/American Stroke Association
CMS	Centers for Medicare & Medicaid Services
EMS	Emergency Medical Services
FY	Fiscal Year
GWTG-Stroke	Get With The Guidelines – Stroke
Hospital IQR	Hospital Inpatient Quality Reporting
HLM	Hierarchical Logistic Regression Model
HVBP	Hospital Value-Based Purchasing
ICC	Intraclass Correlation
IQR	Interquartile Range
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
MSR	Michigan Stroke Registry
NIHSS	National Institutes of Health Stroke Scale
NQF	National Quality Forum
OR	Odds Ratio
O/E	Observed/Expected
P/E	Predicted/Expected
P4P	Pay-for-Performance

PVP	Predictive Value Positive
PVN	Predictive Value Negative
RSMR	Risk-Standardized Mortality Rate
RSRR	Risk-Standardized Readmission Rate
SD	Standard Deviation
Se	Sensitivity
Sp	Specificity

CHAPTER 1: BACKGROUND AND OBJECTIVES

Burden of Stroke in the US

Stroke is the 4th leading cause of death and the leading cause of serious long-term disability in the United States.[1] Recent estimates indicate that there are 795,000 new and recurrent strokes annually[1], with direct medical costs of $17.5 billion in 2011.[2] There are over 1 million hospital admissions for stroke in the US every year. The average inpatient stay for stroke patients is about 6 days in the US[1,3], with the average hospitalization resulting in an estimated $46,518 in charges.[3] Consequently, stroke is the 10th most expensive condition billed to Medicare and Medicaid and private insurers, and the 5th most expensive condition for uninsured patients in the US.[4]

CMS and Pay-for-Performance

In light of this extraordinary clinical and financial burden, the Centers for Medicare & Medicaid Services (CMS) has decided to incorporate a 30-day ischemic stroke risk-standardized mortality rate (RSMR) and readmission rate (RSRR) into its Hospital Inpatient Quality Reporting (Hospital IQR)[5] and Hospital Value-Based Purchasing (HVBP)[6] programs. These programs illustrate the implementation of pay-for-performance (P4P) models in healthcare. P4P models tie provider reimbursement to reporting and predetermined performance measure standards, as opposed to the volume and complexity of services provided in the traditional fee-for-service model of reimbursement.[7] With health expenditures reaching $2.7 trillion in 2011[8] and expected to grow to almost 20% of the US gross domestic product by 2023[9], both private and public healthcare providers are implementing P4P models in an attempt to improve the efficiency of healthcare delivery.[10] The overall mission of the CMS P4P programs is to promote

high-quality, patient-centered care and accountability through the reporting of predetermined performance measures.[11]

The CMS Hospital IQR program was mandated by the Medicare Prescription Drug, Improvement, and Modernization Act in 2003. The program is designed to incentivize hospitals to report on condition-specific quality measures,[5] which are publicly available through the Hospital Compare website. The Hospital Compare program allows health care consumers to find and compare hospitals based on their reported measures.[12] Recently, the Affordable Care Act (ACA) utilized the Hospital IQR infrastructure to tie the reported performance on quality measures to proportional financial reimbursements through the HVBP program.[6] Changes in reimbursement are dictated by adjustment factors, which are determined by a total performance score, which reflecting a combination of clinical processes, patient experience, outcomes, and efficiency of care measures.[13,14] Hospitals put a percent of their reimbursements (currently 1.5%) into a pool, and based on their performance score rank order, either earn back or lose a proportion of that amount.

Additionally, in June 2007, CMS began publicly reporting hospital 30-day RSMRs for acute myocardial infarction (AMI) and heart failure (HF), and subsequently added a 30-day mortality rate for pneumonia in June 2008.[15,16] Hospital 30-day readmission rates (RSRRs) were added for the same conditions in June 2009 as a part of the Hospital Readmissions Reduction Program (HRRP).[17] In 2014, hospitals began to submit 30-day ischemic stroke and chronic obstructive pulmonary disorder (COPD) RSMRs and RSRRs, in addition to the AMI, HF, and pneumonia measures.[16,18] Measures related to clinical processes, patient experience, patient safety, and spending per beneficiary are also publicly reported.

Hospital Profiling and Risk Adjustment

The HVBP program, and more generally, the P4P model presupposes that hospitals can be accurately compared based on predetermined performance measures. The process of comparing hospitals through rank-ordering performance measures (e.g. process rates, outcome rates) is commonly referred to as hospital profiling.[19,20] A critical aspect of hospital profiling is accounting for the variation in patient characteristics between hospitals – referred to as case-mix – using risk-adjustment methods.[21-24] Because patients are not randomized into hospitals, we must use statistical adjustment to account for imbalances in hospital case-mix.[25] Thus, the purpose of risk-adjustment, or case-mix adjustment, is to control for confounding that exists due to differences in the case-mix of patients between hospitals.[19] An important aspect in building risk adjustment models to accurately rank hospitals is including predictors of the outcome that vary between hospitals. If predictors are evenly distributed between hospitals, their inclusion in risk adjustment models will have little effect on improving the accuracy of hospital rankings.[26] The adequacy of risk adjustment is often a focal point of debate; and without satisfactory risk-adjustment, the use of hospital profiling becomes problematic.

All risk adjustment models assume that after accounting for case-mix differences, the resulting differences in hospital outcomes (e.g. RSMR and RSRRs) are due to underlying differences in quality between hospitals that are under control of the hospital.[18] To account for case-mix differences CMS currently uses hierarchical logistic regression modeling (HLM) to calculate a hospital RSMR or RSRR, adjusting for patient case-mix.[19] HLM is a multilevel modeling approach that accounts for the clustering of observations by hospital, and can estimate hospital-specific deviation in an outcome from the population average based on the

estimated hospital random intercept.[20,27,28] This method is generally preferred to indirect standardization by way of standard logistic regression models, as it has been shown to be less sensitive to smaller hospitals that have fewer observed outcome events, avoids regression-to-the-mean bias, and calculate more accurate predicted probabilities based on hospital-level effects.[28-31]

The HLM approach estimates a hospital RSMR, which is calculated as the ratio of "predicted" deaths to "expected" deaths multiplied by the overall mortality rate. The "predicted" number of deaths is the sum of individual predicted probabilities from the multivariable HLM for all patients seen at a particular hospital (which accounts for case-mix), conditional on the hospital's performance on mortality, i.e. the hospital-specific random intercept.[19,32] The "expected" number of deaths is the sum of individual predicted probabilities of death based on case mix, conditional on the average hospital performance, i.e. setting the hospital-specific random intercept to zero. [19,32] The "predicted" to "expected" ratio (P/E ratio) is therefore the ratio of deaths expected at a given hospital compared to the number of deaths expected at the average hospital with the same case-mix. The P/E ratio is then multiplied by the overall mortality rate to get the RSMR. If the "predicted" number of deaths in a hospital is higher than the "expected" number of deaths (i.e. P/E ratio > 1) the resulting RSMR for that hospital would be greater than the overall average mortality rate. Conversely, if there are fewer "predicted" deaths than "expected" deaths, the hospital RSMR would be lower than average.

In addition to calculating RSMRs, the HLM approach can be used to identify statistical outliers in hospital performance using the estimated hospital random intercept. The

distribution of hospital random intercepts is assumed to be a normal distribution centered on zero. Thus, hospitals can be identified as "outlier" hospitals, or hospitals with extreme performance (high or low), based on where the estimated hospital-specific random intercept lies on the normal distribution of random intercepts. Typically, if a hospital random intercept 95% confidence interval does not include 0 (i.e. the hospital average), it is considered an outlier hospital.[33,34] This method has been shown to identify outlier hospitals more accurately compared to the partitioning of hospitals into categories based on their performance measure, such as quintiles of performance, where many hospitals in the lowest or highest quintiles are not statistically identified as outliers.[35]

Hospital-Specific Mortality as a Performance Measure

Despite advances in the statistical methodology used to profile hospitals, a contentious debate surrounds the use of mortality to compare hospitals. Supporters of mortality as a performance measure often cite that mortality is a single, easily interpreted, and clinically meaningful measure to many different stakeholders, especially to patients.[36] They also claim that mortality may reflect an aggregate measures of quality that may not otherwise be identified through other specific quality measures that reflect processes or structural measures.[22] Furthermore, all-cause mortality is considered a highly reliable, universally available, and unambiguous measure across all settings, which makes it an ideal reporting measure.[22] A recent study by McCrum, et al. showed that 30-day RSMRs for AMI, HF, and pneumonia were highly predictive of mortality rates for other medical and surgical conditions within a hospital, suggesting that they may be useful surrogates for overall hospital mortality performance.[37]

However, another study by Jha, et al. found that performances on AMI, HF, and pneumonia mortality rates are not well correlated within a hospital, signifying that overall mortality performance may not adequately identify "good" or "bad" performing hospitals.[38] Other significant limitations with using mortality as a comparative measure of hospital performance include its inability to discriminate well between high and low performing hospitals, the significant impact of coding and risk adjustment methods on resulting measure estimates, the ability for interventions to impact hospital mortality, and that it may be misleading true quality of a hospital. A study by Mackenzie, et al. suggests that RSMR estimates are not precise enough to sufficiently discriminate "good" from "bad" hospitals when used to profiling hospitals.[39] Differences in coding and admission practices across hospitals may also bias hospital standardized mortality ratios, which may incorrectly attribute differences in outcomes between hospitals to underlying differences in quality of care.[40] The methods by which RSMRs are risk adjusted have also been shown to produce substantially different results, even though they were applied to the same population.[41-43] A recent review of conceptual and methodological challenges of hospital-wide mortality measures concluded that while mortality rates may provide useful information, they may also obscure or distort important signals of quality that are of interest to various stakeholders.[36] Importantly, Hogan, et al. found that while mortality is a clinically relevant measure, few hospital deaths are preventable, which would limit its value as an endpoint for quality improvement initiatives aimed to improve hospital performance.[44] A study of mortality following coronary artery bypass graft surgery showed that only one third of in-hospital deaths were deemed preventable.[45] Nonetheless, while debate rages about the appropriateness of using of hospital-wide mortality as a

performance measure to compare hospitals, public and private payers are forging ahead and incorporating them into their P4P programs.

Controversy with 30-Day Ischemic Stroke Measures

The recent addition of the 30-day ischemic stroke RSMR and RSRR to the Hospital IQR and HVBP programs has been especially contentious. Currently, they lack support from the National Quality Forum – a non-partisan organization which evaluates proposed performance measures – and the American Heart Association/American Stroke Association (AHA/ASA).[46-48] The primary reason cited for opposing the RSMR and RSRR measures is that they are inadequately risk-adjusted due to the exclusion of a measure of stroke severity, such as the National Institute of Health Stroke Scale (NIHSS).[47-49] The NIHSS[50] is a commonly used measure of stroke severity collected in stroke trials and registries[51], which includes functional domains of level of consciousness, horizontal eye movement, visual field test, facial palsy, arm motor function, leg motor function, limb ataxia, sensory perception, language impairment, and speech impairment.[50] (Table 1.1) In a Presidential Advisory statement from the AHA/ASA, Fonarow et al. state that the "outcome measures as currently constructed may be prone to mischaracterizing the quality of stroke care being delivered by hospitals and may ultimately harm ischemic stroke patients."[48]

Table 1.1. Domains and score/descriptions for National Institute of Health Stroke Scale, final score ranges from 0-42.

Domain		***Score/Description***
1a.	Level of Consciousness (Alert, drowsy, etc.)	0 = Alert 1 = Drowsy 2 = Stuporous 3 = Coma
1b.	LOC Questions (Month, age)	0 = Answers both correctly 1 = Answers one correctly 2 = Incorrect
1c.	LOC Commands (Open/close eyes, make fist let go)	0 = Obeys both correctly 1 = Obeys one correctly 2 = Incorrect
2.	Best Gaze (Eyes Open – Patient follows examiners finger or face)	0 = Normal 1 = Partial gaze palsy 2 = Forced deviation
3.	Visual Fields (Introduce visual stimulus/threat to patients visual field quadrants	0 = No visual loss 1 = Partial hemianopia 2 = Complete hemianopia 3 = Bilateral hemianopia (blind)
4.	Facial Paresis (Show teeth, raise eyebrows and squeeze eyes shut)	0 = Normal 1 = Minor 2 = Partial 3 = Complete
5a. 5b.	Motor Arm - Left Motor Arm – Right (Elevate arm to 90^{o} with patient supine)	0 = No drift 1 = Drift 2 = Can't resist gravity 3 = No effort against gravity 4 = No movement X = Untestable (Joint fusion or limb amputation)
6a. 6b.	Motor Leg – Left Motor Leg – Right (Elevate leg to 30^{o} with patient supine)	0 = No drift 1 = Drift 2 = Can't resist gravity 3 = No effort against gravity 4 = No movement X = Untestable (Joint fusion or limb amputation)
7.	Limb Ataxia (Finger-nose, heel down shin)	0 = No ataxia 1 = Present in one limb 2 = Present in two limbs
8.	Sensory (Pin prick to face, arm, trunk and leg – compare side to side)	0 = Normal 1 = Partial loss 2 = Severe loss
9.	Best Language (Name item, describe a picture, and read sentences)	0 = No aphasia 1 = Mild to moderate aphasia 2 = Severe aphasia 3 = Mute
10.	Dysarthria (Evaluate speech clarity by repeating listed words)	0 = Normal articulation 1 = Mild to moderate alluring of words 2 = Near to unintelligible or worse X = Intubated or other physical barrier
11.	Extinction and Inattention (Use information from prior testing to identify neglect or double simultaneous stimuli testing)	0 = No neglect 1 = Partial neglect 2 = Complete neglect

As is done with current 30-day RSMRs for AMI, HF, and pneumonia, CMS administrative data are used to generate RSMR and RSMRR used to determine hospital-level performance.[23,24,35,52] Absent from CMS administrative claims data is a measure of stroke severity, such as the NIHSS. Studies have shown that measures of stroke severity, such as the NIHSS, significantly improve prediction of patient-level stroke outcomes and are widely believed to be essential for risk adjustment at the hospital-level.[53-56] A systematic review of case-mix adjustment models for post-stroke mortality and functionality found that stroke severity is a commonly used and important variable in individual-level risk-adjustment.[57] However, it is unclear if stroke severity varies substantially across hospitals enough to make it a significant confounder. A study of Veterans Affairs (VA) hospitals showed that the addition of NIHSS into risk adjustment had minimal improvement on model fit, most likely due to little variation in NIHSS between VA hospitals.[58] There are little data on the true variation in stroke severity across all US hospitals.

Due to its importance in individual-level prediction models, there is concern that excluding stroke severity from risk adjustment will lead to incorrect rankings of hospital performance, particularly in stroke referral centers that typically see a more severe spectrum of patients.[26,48,59] In a similar situation, a study conducted by Friese et al. that compared outcomes in surgical cancer patients, the severity of cancer varied significantly between hospitals, and resulting risk-adjusted hospital mortality rates were lower among hospitals with less severe patients compared to hospitals with more advanced disease patients when cancer severity was not included in the risk adjustment model.[60] Furthermore, studies of ICU performance have shown that referral centers – which frequently accept more severe patients

– typically have higher RSMRs compared to referring centers.[61,62] Therefore, it is reasonable to believe that not accounting for stroke severity in risk adjustment may similarly bias hospital ischemic stroke RSMRs, assuming that there is significant variation in stroke severity between hospitals.

Current Limitations to Including NIHSS in Risk Adjustment Models

To date, there has been conflicting evidence supporting the use of NIHSS in risk adjustment for hospital profiling. One study conducted by Fonarow, et al. found that among hospitals profiled into the top or bottom 20% according to their RSMRs, 26% were ranked differently once NIHSS was included in risk adjustment.[53] However, this was in a dataset with >50% missing NIHSS data. As previously mentioned, data from VA hospitals showed little variation in NIHSS between hospitals, and hospital RSMRs calculated with and without NIHSS were nearly identical.[58] It is yet unclear if there is sufficient variation in stroke severity between hospitals – a necessary condition for risk-adjustment variables[26] – especially among hospitals that are assumed to treat a more severe set of patients, such as tertiary referral centers and certified primary stroke centers.[63]

A more practical limitation to including stroke severity in risk adjustment cannot be ignored. Unlike CMS administrative claims data, clinical registries often do collect measures of stroke severity. But, despite recent improvements in documentation, registries still struggle to achieve complete reporting of stroke severity.[53,54,56] Given that hospital-specific measures are calculated from risk-adjustment models using only cases with complete data on risk adjustment variables, i.e. a complete case analysis, resulting measures can be significantly biased when a biased subset of patients are used.[64,65] One study suggests that assessments of mortality using

a complete case analysis of subjects with observed NIHSS may be subject to bias in hospitals with very low documentation of NIHSS.[66] Unless complete reporting of NIHSS can be achieved through CMS administrative data, hospital-wide measures calculated from incomplete data may be biased.

Bias as a Result of Missing Data

The extent of bias from a complete case analysis of incomplete data depends on the mechanism by which data are missing. Missing data are typically classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).[67] If data are MCAR, the missing data are unassociated with any exposure or outcome information. In other words, missing data are the result of a purely random incident, and the observed data are a random sample of the entire data. In theory, a complete case analysis under MCAR should result only in a loss of statistical efficiency (because of the smaller sample size), but not produce biased estimates.[68] If data are MAR, missing data are associated with fully observed variables. For example, if stroke severity documentation is better in males compared to females (all observed), data would be considered MAR. In addition to a loss in statistical efficiency, complete case analysis under MAR may result in biased estimates if the reason for missing data (gender) is not accounted for.[68,69] Methods such as maximum likelihood estimation and multiple imputation can be employed to combat biased estimates and a loss in statistical efficiency when data are MCAR or MAR.[68,69]

The most problematic missing data scenario is when data are MNAR, which is to say that missingness is related to either unobserved characteristics or the value of the missing variable itself.[68] For example, if stroke severity documentation was better in patients with more severe

strokes compared to less severe strokes, the data would be MNAR. Again, bias and loss of statistical efficiency are attributed to MNAR data. However, the methods employed when data are MCAR or MAR cannot correct for all the bias resulting from MNAR data, because you cannot directly estimate a pattern based on missing data.[68,69]

Missing data are common in clinical research.[68,70] They are especially common in administrative datasets such as billing data, where certain variables may be completely unavailable, or data from electronic health records, where variables are often incompletely documented.[71] Research has shown that a complete case analysis when covariate data are missing can lead to biased estimates of patient-level outcomes.[72-74] How a complete case analysis in the presence of missing data impacts hospital-level estimates is less obvious. One simulation study comparing hospital trauma-related mortality measures showed that a complete case analysis when risk-adjustment variable data are MNAR led to considerable changes in hospital-level mortality profiling.[64] Using a complete case analysis to profile hospitals when missing data are present has also been shown to underestimate the proportion of poorly performing providers.[75] Another simulation study examining the impact of missing data on profiling in P4P outcomes showed that between 11 to 21 percent of misclassified hospitals were attributable to missing data in risk adjustment.[65]

An analogous problem to excluding patients in hospital-level measures based on incomplete documentation is variation in administrative data coding. While there are already well documented limitations to using administrative data in hospital profiling[71,76-80], differences in coding of data can lead to differential exclusion of patients between hospitals. There are a number of examples that illustrate how variations in coding between hospitals impact hospital-

level measures. An analysis of data in the United Kingdom showed that differential coding of comorbidities between hospitals case-mix adjustment may create biased hospital RSMRs.[40] Another recent study demonstrated that excluding patients from pneumonia RSMR calculations due to variation in coding for pneumonia misclassified 28% of hospitals.[81] Austin, et al. suggested that undercoding of significant comorbidities or severity indicators, which makes patients appear healthier than they actually are, can potentially misclassify hospitals.[82] Using a "present-on-admission" indicator to distinguish between existing comorbidities and complications related to quality of care when risk-adjusting for patient health status showed that a quarter of hospital AMI mortality rankings were misclassified by 10% or more.[83]

In sum, there is a multitude of research showing that excluding patients from hospital-level measures, either due to missing clinical data or administrative coding variation, can lead to inaccurate hospital profiling. However, it is unclear how different mechanisms of missing data, and the frequency at which missing data occur, can impact the accuracy of hospital profiling. To our knowledge, this is the first study to assess how different mechanisms and frequencies of missing NIHSS data impact the accuracy of hospital profiling of stroke mortality measures.

Statement of Problem, Aims, and Outline

Currently, administrative data used to profile hospitals on CMS 30-day ischemic stroke RSMRs do not collect measures of stroke severity, such as NIHSS. When NIHSS is collected in clinical data, such as stroke registries, it is frequently missing and little is known about what predicts NIHSS documentation. If NIHSS is to be included in risk-adjustment models, cases with missing NIHSS will be excluded in the calculation of the hospital 30-day RSMR for ischemic

stroke. The resulting RSMR may be biased, depending on the mechanism and frequency of missing NIHSS data. Ultimately, biased RSMRs could lead to inaccurate hospital profiling, which may unfairly distribute financial incentives in P4P reimbursement models. The aims of this study are as follows:

1) To identify significant patterns or predictors of NIHSS documentation at the patient-level and hospital-level in an existing stroke registry.
2) To test for the presence and magnitude of selection bias in patients with documented NIHSS using the Heckman Selection Model.
3) To estimate the impact of the prevalence and mechanism of missing NIHSS data on the accuracy of hospital profiling of 30-day ischemic stroke RSMRs using computer simulation models.

The subsequent chapters of this dissertation will be organized by answering questions for each of these aims.

What are the overall patterns or predictors of patient-level NIHSS documentation at the patient- and hospital-level? Chapter 2 will test the hypothesis that there are significant patient and hospital predictors of NIHSS documentation. Using data from the Michigan Stroke Registry, we will provide insight into patient or hospital characteristics that explain the documentation of NIHSS data in stroke patients. Analyses of NIHSS documentation to identify patterns and predictors will help identify the mechanism and pattern of missing NIHSS data.

Is the subset of patients with NIHSS documented a biased sample, and, if so, to what extent? Chapter 3 will assess the presence of selection bias in the documentation of NIHSS using the Heckman Selection Method. The Heckman Selection Method will be used as a

diagnostic test for the presence of selection bias in patients with NIHSS documented, i.e., patients with observed NIHSS data are systematically different from patients with unobserved NIHSS. While the previous aim helps to identify significant patterns and predictors of NIHSS documentation, this aim will provide statistical evidence for selection bias in NIHSS documentation based on patient stroke severity. The Heckman model also indicates the magnitude and direction of selection bias in patients with undocumented NIHSS data. Furthermore, if there is significant selection bias, it would suggest that missing NIHSS data are MNAR, or non-ignorable. Jointly, the first and second aims will provide a clearer picture of the mechanism and pattern of NIHSS documentation, which will motivate the use of different missing data mechanisms in the subsequent computer simulations used in Aim 3.

How does the presence of missing data impact the accuracy of hospital performance profiling? What role does the prevalence and mechanism of missing NIHSS data have on the accuracy of hospital profiling? How does hospital case volume modify this relationship? Chapter 4 will assess the hypothesis that the accuracy of hospital profiling will be affected in datasets with missing NIHSS compared to fully documented data. This aim will illustrate how sensitive hospital profiling is when RSMRs are calculated in the face of missing data. Furthermore, it will illustrate which mechanisms and patterns of NIHSS documentation result in the most inaccurate hospital rankings at various frequencies of NIHSS documentation. Finally, we will assess how missing data impact profiling at different hospital ischemic stroke case volumes.

These analyses will demonstrate the accuracy of hospital profiling based on risk-adjusted mortality models when an important risk adjustment variable is frequently

undocumented. Furthermore, it will illustrate what role the mechanism of missing data plays in profiling accuracy. Finally, it will illustrate the importance of hospital volume as an vital modifier of the relationship between missing data and profiling accuracy. As Voltaire is famously quoted, “It is better to risk saving a guilty person than to condemn an innocent one.” We seek to quantify just how many guilty hospitals are saved, and more importantly, how many innocent hospitals will be condemned.

CHAPTER 2: PATTERNS AND PREDICTORS OF NIHSS DOCUMENTATION

Aim 1 – Background

The National Institutes of Health Stroke Scale (NIHSS)[50] is a commonly used measure of stroke severity collected in stroke trials and registries.[51] NIHSS has been shown to be one of the strongest predictors of outcomes in ischemic stroke patients.[54,56,84] Despite its clinical importance, complete documentation of NIHSS in clinical registries has yet to be achieved. While NIHSS documentation has improved recently, documentation was below 50% in the first 5 years of the Get With the Guidelines (GWTG) – Stroke national registry.[54] Furthermore, measures of stroke severity are currently absent from administrative data.

The Centers for Medicare & Medicaid Services (CMS) will soon be adding 30-day measures of hospital-level ischemic stroke mortality and readmissions to its pay-for-performance incentive programs.[5,6] Because of its importance as a clinical prognostic variable at the patient-level, it is believed that risk adjustment models used to calculate hospital-level performance metrics must include a measure of stroke severity,[48,59] although evidence to support this claim has been mixed.[53,58] Given that complete documentation of NIHSS has not been achieved, excluding patients with undocumented NIHSS from risk adjustment models may impact the validity of hospital-level performance measures. Furthermore, any bias in hospital-level measures may be aggravated if NIHSS data are missing not at random (MNAR).[67,68]

Because variation in NIHSS documentation has the potential to bias hospital-level ischemic stroke performance measures, the purpose of this study is to describe trends in NIHSS documentation in an existing multi-center clinical stroke registry, and identify any significant patient- and hospital-level factors associated with NIHSS documentation. Also, we will attempt

to determine the extent of bias in NIHSS scores by describing the relationship between NIHSS documentation and NIHSS scores at the hospital-level. In essence, we will try to determine the mechanism by which NIHSS is missing. We hypothesize that patients with documented NIHSS are not simply a random sample of all patients, and that missing NIHSS data may be MNAR.

Aim 1 – Methods

Data and Participants

The Michigan Stroke Registry (MSR) is a statewide clinical registry which originated as a prototype for the Paul Coverdell National Acute Stroke Registry, and has been described elsewhere.[85] Currently, the MSR is used to provide a data driven approach to improve the quality of stroke care in the State of Michigan.[86,87] The MSR collects information on many different patient level characteristics including demographics, emergency medical services (EMS) and hospital admission information, and clinical information such as stroke severity, ambulatory status, and medical history. In addition, we obtained hospital characteristics from the American Hospital Association annual survey[88] and Paul Coverdell National Acute Stroke Registry hospital inventory.

We used MSR data from 2009 to 2012 for this analysis. To increase the generalizability of our findings to a CMS ischemic stroke population, we applied a number of exclusions to the MSR data. Ischemic stroke patients were included if they were aged 65 years or older and excluded if they belonged to a hospital with <25 annual cases of ischemic stroke, which is the minimum number of cases for a hospital risk-standardized mortality rate (RSMR) to be calculated, as defined by CMS.[18] We also excluded patients if the stroke occurred in a hospital inpatient setting. As this study was a secondary analysis of deidentified registry data, it was

considered exempt from Institutional Review Board review. All analyses were conducted with the use of SAS version 9.3 (SAS Institute Inc, Cary, NC).

Predictor Variables

We examined a number of patient-level predictors of NIHSS documentation. Demographic characteristics included: age, gender (male vs. female), race (white, black, other, not documented), and insurance status (Medicare, Medicaid, private, no insurance). We also assessed emergency medical services (EMS) and hospital admission information, such as: place stroke occurred (at home vs. in a healthcare setting), arrival mode (EMS, private transportation, transferred), arrival to the ER (yes vs. no), symptoms resolved prior to arrival (yes vs. no), and tPA administration (yes vs. no). Finally, we also examined several clinical variables in this analysis, including: able to ambulate pre-stroke, diabetes mellitus, congestive heart failure, peripheral artery disease, hypertension, current smoker, and history of prior stroke, transient ischemic attack/vertebrobasilar insufficiency (TIA/VBI), or myocardial infarction/coronary artery disease (MI/CAD).

Hospital characteristics included bed size, annual stroke volume (<200, 200-600, 600+), urban vs. rural location, teaching status, presence of an acute stroke team, and Joint Commission primary stroke center status.[63]

Outcome Variable

The NIHSS is a composite measurement of eleven symptoms measurements, including level of consciousness, horizontal eye movement, visual field test, facial palsy, arm motor function, leg motor function, limb ataxia, sensory perception, language impairment, and speech impairment.[50] (Table 1.1) The resulting score is an integer which ranges from 0 to 42, with 0

representing no stroke symptoms and 42 representing the most severe form of stroke. In patient-level analyses, we used a binary NIHSS documentation indicator (yes vs. no) as the outcome variable. In hospital-level analyses, we used the patient- and hospital-level average NIHSS score as the outcome variable. Hospital-level NIHSS documentation rates were calculated and categorized by tertiles of NIHSS documentation (<70%, 70-85%, and ≥85%) to represent low, moderate, and high documenting hospitals.

Statistical Analysis

First, a patient-level descriptive analysis of the data was conducted, which assessed the distribution of demographic, EMS and hospital admission information, and clinical variables, as well as patient-level hospital characteristics in the sample, stratified by NIHSS documentation (yes vs. no). To identify patient-level factors associated with documentation, bivariate associations were assessed using chi-square tests and ANOVA for categorical and continuous variables, respectively. We also assessed differences in hospital characteristics, mean NIHSS score, mortality rates and average length of stay (in days) between tertiles of NIHSS documentation rate. Fisher's Exact Test and ANOVA were used to test for any significant differences between tertile for categorical and continuous variables, respectively. We then tested for significant changes in hospital-level documentation rates over time with ANOVA, which were then illustrated using box plots for each year (2009-2012).

Significant patient- and hospital-level predictors of NIHSS documentation at the patient-level were assessed using unadjusted and adjusted hierarchical logistic regression model, which accounting for clustering of data within hospitals. The modeling procedure was motivated by the multilevel modeling approach by Singer.[89] First, a model with a hospital random intercept

and no fixed effects was run to assess the within-hospital variation in NIHSS documentation. The hospital-level variance (σ_j) was used to calculate the intraclass correlation coefficient (ICC) in the model using the equation, $ICC = \sigma_j / \left(\sigma_j + \frac{\pi^2}{3}\right)$. Then, we specified a fully saturated model, which included all patient and hospital-level variables with p<0.20 in the previous bivariate analysis as fixed effects, as well as a hospital random intercept. Using a backward selection approach with stepwise deletion, we eliminated all non-significant (p>0.05) variables from the model. The final model contained significant patient and hospital fixed effects, and a hospital random intercept. We tested for the statistical significance of σ_j using a log-likelihood test. *A priori* hospital-level fixed effects terms, including primary stroke center status and stroke volume, were retained in the final model regardless of their statistical significance.

To determine if NIHSS documentation is related to the NIHSS score, i.e. undocumented NIHSS is MNAR, we performed two analyses. First, Pearson and Spearman correlation coefficients were calculated to assess relationships between hospital-level NIHSS documentation and hospital-level NIHSS score, for each year (2009-2012). A significant correlation indicates that the level of documentation is associated with the observed NIHSS score, suggesting data may be MNAR. Second, we tested for significant differences in the patient-level distribution of NIHSS scores stratified by tertile of hospital-level NIHSS documentation rate using ANOVA and a Kruskal-Wallis test. Differences in patient-level NIHSS score distributions by tertile of hospital documentation rate were illustrated by overlaying smoothed frequency distributions (kernel density curves) for patients within each tertile. A shift in the distribution in lower levels of hospital-level documentation may also suggest data may be MNAR.

Aim 1 – Results

Between 2009 and 2012, 18,280 ischemic strokes admitted to 39 hospitals were abstracted from the Michigan Stroke Registry. A total of 6,572 cases (36.0%) were excluded because they were under the age of 65. We also excluded data from 16 hospitals (n=991 cases, 5.4%) because their annual case load was below 25 cases.[18] Therefore, the final sample contained 10,717 cases from 23 hospitals, of which 7,956 cases (74.2%) had NIHSS documented. (Table 2.1) The mean (standard deviation=SD) and median (interquartile range= IQR) for patients with documented NIHSS was 7.3 (SD=7.8) and 4 (IQR=2-11) respectively.

Table 2.1 shows the patient demographics, EMS and hospital admission information, and clinical information of the sample, stratified by NIHSS documentation status. Patients with NIHSS documented were more likely to be white compared to patients who did not have NIHSS documented (74.1% vs. 68.3%). Patients with NIHSS documented were less likely to have Medicaid (4.8% vs. 6.6%) and more likely to be privately insured (45.5% vs. 43.2%) compared to those with NIHSS undocumented. Patients who had NIHSS documented also tended to be at home at the time of onset (90.6% vs. 86.5%, $p<0.0001$) and were more likely to arrive to the ER (89.1% vs. 85.9%, $p<0.0001$). (Table 2.1) There was a marked and significantly lower percent of patients whose symptoms had resolved by hospital arrival in those with NIHSS documented compared to undocumented (3.6% vs. 15.6%, $p<0.0001$). Another striking difference was in tPA administration rates between those with and without NIHSS documented (9.3% vs. 1.0%, $p<0.0001$). (Table 2.1) In regard to patient medical history, patients with NIHSS documented were slightly more likely to be ambulatory pre-stroke (96.0% vs. 92.6%, $p<0.0001$), had higher rates of atrial fibrillation (23.7% vs. 19.9%, $p<0.0001$) and dyslipidemia (47.3% vs. 40.4%,

p<0.0001), and a lower rate of prior stroke (26.8% vs. 30.5%, p=0.0002). (Table 2.1) There were no significant differences in NIHSS documentation by age, gender, mode of arrival, history of diabetes mellitus, prior TIA/VBI, MI/CAD, congestive heart failure, peripheral artery disease, hypertension, or smoking status.

There were a number of patient-level hospital characteristics which were associated with NIHSS documentation. (Table 2.1) Patients with NIHSS documented tended to be treated at hospitals with slightly fewer beds (p<0.0001) and fewer stroke discharges (p<0.0001). (Table 2.1) Modest differences in the proportion of patients with NIHSS documented were observed between hospitals which were rural vs. urban hospitals, teaching vs. non-teaching hospitals, possessed an acute stroke team, and primary stroke center status as certified by the Joint Commission. (Table 2.1)

Table 2.1. Patient demographics, EMS and admission information, medical history and discharge status in Ischemic Stroke patients 65 years of age or older, in the overall sample (n=10,717) and stratified by NIHSS Documentation status. (2009-2012)

Variable	***NIHSS Documentation Status***		***p-value***
	Documented	*Undocumented*	
	n (%)	*n (%)*	
Overall Sample	7,957 (74.3)	2,760 (25.8)	
Demographics			
Age, *mean (SD)*	78.6 (8.2)	78.8 (8.4)	0.2975
Female	4,345 (54.6)	1,512 (54.8)	0.8773
Race	-	-	<0.0001
White	5,897 (74.1)	1,884 (68.3)	
Black	1,295 (16.3)	646 (23.4)	
Other	93 (1.2)	23 (0.8)	
Not Documented	672 (8.5)	207 (7.5)	
Insurance status	-	-	0.0010
Medicare	3,884 (48.9)	1,364 (49.6)	
Medicaid	378 (4.8)	181 (6.6)	
Private	3,619 (43.2)	1,189 (43.2)	
None	61 (0.8)	17 (0.6)	
EMS and Admission			
Place stroke occurred	-	-	<0.0001

Table 2.1. (cont'd) Patient demographics, EMS and admission information, medical history and discharge status in Ischemic Stroke patients 65 years of age or older, in the overall sample (n=10,717) and stratified by NIHSS Documentation status. (2009-2012)

Variable	***NIHSS Documentation Status***		***p-value***
	Documented	*Undocumented*	
At home	7,208 (90.6)	2,386 (86.5)	
In a healthcare setting	755 (9.4)	375 (13.4)	
Arrival Mode	-	-	0.0829
EMS	3,892 (49.9)	1,310 (48.1)	
Private	2,673 (34.3)	996 (36.7)	
Transfer	1,230 (15.8)	411 (15.1)	
Arrived in the ER	7,092 (89.1)	2,372 (85.9)	<0.0001
Symptoms resolved	279 (3.6)	414 (15.6)	<0.0001
tPA Administration	743 (9.3)	27 (1.0)	<0.0001
Medical History			
Ambulatory Pre-Stroke	7,957 (96.0)	2,368 (92.6)	<0.0001
Atrial Fibrillation	1,882 (23.7)	548 (19.9)	<0.0001
Diabetes Mellitus	2,612 (32.8)	947 (34.3)	0.1534
Prior Stroke	2,130 (26.8)	842 (30.5)	0.0002
Prior TIA/VBI	904 (11.4)	306 (11.0)	0.6198
MI or CAD	2,704 (34.0)	929 (33.7)	0.7572
CHF	1,050 (13.2)	395 (14.3)	0.1392
Peripheral Artery Disease	504 (6.3)	180 (6.5)	0.7281
Dyslipidemia	3,766 (47.3)	1,114 (40.4)	<0.0001
Hypertension	6,442 (81.1)	2,244 (81.3)	0.6910
Smoking	882 (11.1)	342 (12.4)	0.0629
Hospital Characteristics			
Bed size, *Mean (SD)*	505.1 (245.1)	578.6 (274.3)	<0.0001
Acute stroke discharges	-	-	<0.0001
<200	758 (9.5)	211 (7.6)	
200-600	3,339 (42.0)	801 (29.0)	
600+	3,860 (48.5)	1,748 (63.3)	
Mean (SD)	569.3 (308.3)	671.9 (373.7)	<0.0001
Rural Hospital	1,223 (15.9)	290 (11.5)	<0.0001
Teaching Hospital	7,347 (92.3)	2,605 (94.8)	0.0003
Acute stroke team	6,713 (84.4)	2,426 (87.9)	<0.0001
Joint Commission Primary Stroke Center	6,257 (78.6)	2,092 (75.8)	0.0020

Note: Categories with small n were excluded from the table, so cells may not add up to 100%.

Hospital-level characteristics stratified by tertile of hospital NIHSS documentation rate (<70%, 70-85%, ≥85%) can be seen in Table 2.2. Median documentation rates in low, moderate, and high documenting hospitals was 52.7%, 74.8%, and 89.1%, respectively. (Table 2.2) Mean NIHSS scores were significantly different (p=0.0122) between hospitals with low (mean=8.8), moderate (mean=7.1) and high (mean=6.7) documenting hospitals. (Table 2.2) There were no statistically significant differences between any of the hospital characteristics, tPA administration rates, mortality rates, or average length of stay between tertile of hospital documentation. (Table 2.2) Although non-significant, low and moderate documenting hospitals had greater annual stroke volumes compared to high documenting hospitals (p=0.0703).

Table 2.2. Michigan Stroke Registry hospital-level characteristics in the sample of 23 hospitals, stratified by tertile of hospital NIHSS documentation rate.

Variable	***Hospital NIHSS Documentation Rate***			***p-value***
	1st Tertile: <70% (n=7)	*2nd Tertile: 70-85% (n=8)*	*3rd Tertile: ≥85% (n=8)*	
Num. of Patients	2,663 (24.8)	4,749 (44.3)	3,305 (30.8)	-
Characteristics, n (%)	*n (%)*	*n (%)*	*n (%)*	
*Bed Size**	443 (88-675)	407 (311-546)	390 (211-407.5)	0.5092
Annual Stroke Discharges				0.0703
<200	2 (28.6)	1 (12.5)	2 (25.0)	
200-600	2 (28.6)	2 (25.0)	6 (75.0)	
600+	3 (42.9)	5 (62.5)	0 (0.0)	
Rural Hospital	2 (28.6)	1 (12.5)	2 (25.0)	0.8369
Teaching Hospital	5 (71.4)	8 (100.0)	7 (87.5)	0.2727
Acute Stroke Team	6 (85.7)	6 (75.0)	7 (87.5)	1.0000
Primary Stroke Center	4 (57.1)	7 (87.5)	5 (62.5)	0.5299
tPA Administration Rate (%)†	8.4 (4.9)	6.5 (3.6)	6.3 (3.6)	0.5460
Mortality Rate†	6.2 (3.9)	4.6 (1.3)	4.4 (1.4)	0.3029
Avg. Length of Stay (in days)†	5.9 (1.7)	4.8 (0.8)	4.5 (0.8)	0.1854
Hospital-Level NIHSS				
NIHSS Score†	8.8 (1.6)	7.1 (1.2)	6.7 (1.1)	0.0122
*NIHSS Score**	8.7 (7.0-9.7)	7.6 (5.8-8.0)	6.9 (5.8-7.4)	-
Documentation Rate†	52.7 (14.5)	74.8 (2.7)	89.1 (2.7)	<0.0001

* Median (IQR), † Mean (SD)

As illustrated in Figure 2.1, hospital-level NIHSS documentation rates have significantly improved over time (p=0.0072), from a median of 66.8% (IQR: 52.4-76.3%) in 2009 to 86.8% (IQR: 71.7-92.8%) in 2012.

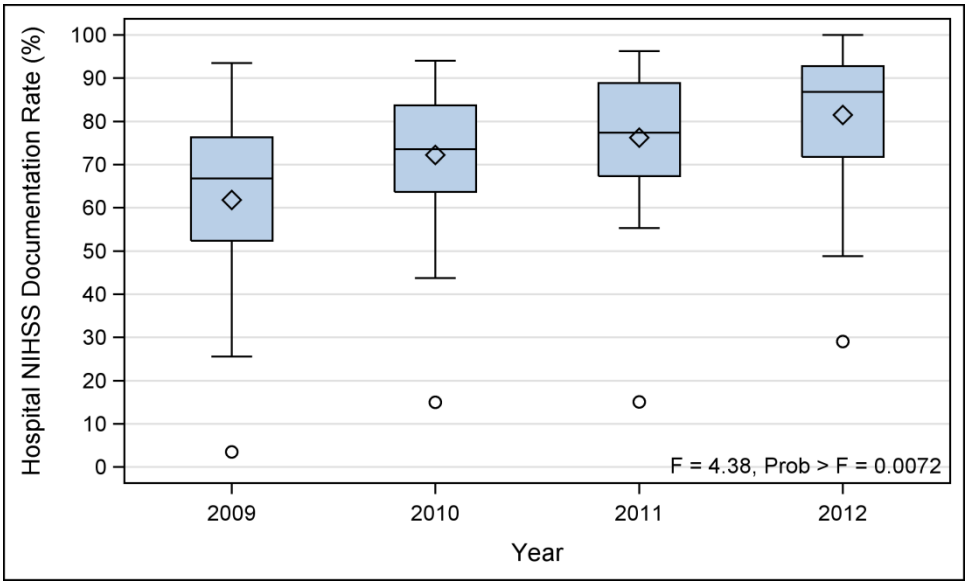

Figure 2.1. Hospital-level NIHSS documentation rates over time.

Table 2.3 presents the results of the final hierarchical logistic regression model fitted to predict patient-level NIHSS documentation (yes vs. no) based on patient- and hospital-level characteristics. After adjustment, patient-level predictors of NIHSS documentation included the stroke occurring at home (OR=1.22; 95% CI: 1.01, 1.48), mode of arrival (hospital transfer vs. private transport OR=1.29; 95% CI: 1.05, 1.58), ER presentation (OR=1.69; 95% CI: 1.36, 2.11), and if the patient was administered tPA (OR=11.46; 95% CI: 7.31, 17.99). NIHSS documentation was also predicted by pre-stroke ambulatory status (OR=1.75; 95% CI: 1.37, 2.23) and a medical history of atrial fibrillation (OR=1.17; 95% CI: 1.02, 1.34) and dyslipidemia (OR=1.15; 95% CI: 1.02, 1.28). Patients with a prior stroke (OR=0.86; 95% CI: 0.76, 0.97) and patients whose symptoms resolved by ER arrival (OR=0.13; 95% CI: 0.11, 0.16) were also at significantly

reduced odds of NIHSS documentation. Although non-significant, large hospitals had a reduced odds of documentation, and Joint Commission primary stroke centers had an increased odds of documentation. We estimated a statistically significant hospital-level variance as $\sigma_j = 1.0930$ (p<0.0001), and calculated the $ICC = 1.0930/\left(1.0930 + \frac{\pi^2}{3}\right) = 0.249$ or 24.9%, which suggests that roughly a quarter of the unexplained variation in NIHSS documentation can be attributed to the hospital-level.

Table 2.3. Unadjusted and adjusted odds ratios (and 95% CIs) for patient and hospital characteristics predicting NIHSS documentation (yes vs. no) and estimated hospital-level variation and intraclass correlation (n=10,717).

Variable	***Unadjusted OR (95% CI)***	***p-value***	***Adjusted OR (95% CI)***	***p-value***
Place Stroke Occurred				
Home	1.51 (1.32, 1.72)	<0.0001	1.22 (1.01, 1.48)	0.0377
Healthcare Setting	Ref	-	Ref	-
Arrival Mode	-	0.0830	-	0.0520
EMS	1.11 (1.01, 1.22)	0.0379	1.08 (0.95, 1.21)	0.2302
Transfer	1.12 (0.98, 1.27)	0.1091	1.29 (1.05, 1.59)	0.0179
Private	Ref	-	Ref	-
Received in ER	1.34 (1.18, 1.53)	<0.0001	1.69 (1.36, 2.11)	<0.0001
Symptoms Resolved	0.20 (0.17, 0.23)	<0.0001	0.13 (0.11, 0.16)	<0.0001
tPA Administered	10.43 (7.08, 15.34)	<0.0001	11.46 (7.31, 17.99)	<0.0001
Ambulatory Pre-Stroke	1.88 (1.56, 2.27)	<0.0001	1.75 (1.37, 2.23)	<0.0001
History of Atrial Fibrillation	1.25 (1.12, 1.39)	<0.0001	1.17 (1.02, 1.34)	0.0296
History of Dyslipidemia	1.33 (1.22, 1.45)	<0.0001	1.15 (1.02, 1.28)	0.0198
History of Prior Stroke	0.83 (0.76, 0.92)	0.0002	0.86 (0.76, 0.97)	0.0151
Year	-	<0.0001	-	<0.0001
2012	2.70 (2.38, 3.09)	<0.0001	3.03 (2.59, 3.55)	<0.0001
2011	1.83 (1.63, 2.06)	<0.0001	1.88 (1.63, 2.17)	<0.0001
2010	1.49 (1.33, 1.68)	<0.0001	1.39 (1.21, 1.61)	<0.0001
2009	Ref	-	Ref	-
Primary Stroke Center	1.18 (1.06, 1.32)	0.0022	1.86 (0.65, 5.30)	0.2320
Stroke Discharges				0.4219
600+	0.62 (0.52, 0.72)	<0.0001	0.53 (0.15, 1.95)	0.3224
200-600	1.16 (0.98, 1.38)	0.0881	1.03 (0.30, 3.47)	0.9660
<200	Ref	-	Ref	-
Estimated hospital-level variance, $\sigma_j = 1.0930$	1.09 (0.37, 1.81)	<0.0001	$ICC = 1.0930/\left(1.0930 + \frac{\pi^2}{3}\right)$ $ICC = 24.9\%$	

ICC = Intraclass correlation

Figure 2.2 plotted the hospital-level documentation rate against the mean hospital-level NIHSS score for all hospitals (n=23) in each year (2009-2012). The significant, negative Pearson ($r = -0.44$, $p<0.0001$) and Spearman ($r = -0.39$, $p<0.0001$) correlation coefficients indicate moderate correlation between hospital-level NIHSS documentation and NIHSS score. This suggests that at the hospital-level, mean observed NIHSS scores were higher amongst hospitals with lower documentation of NIHSS.

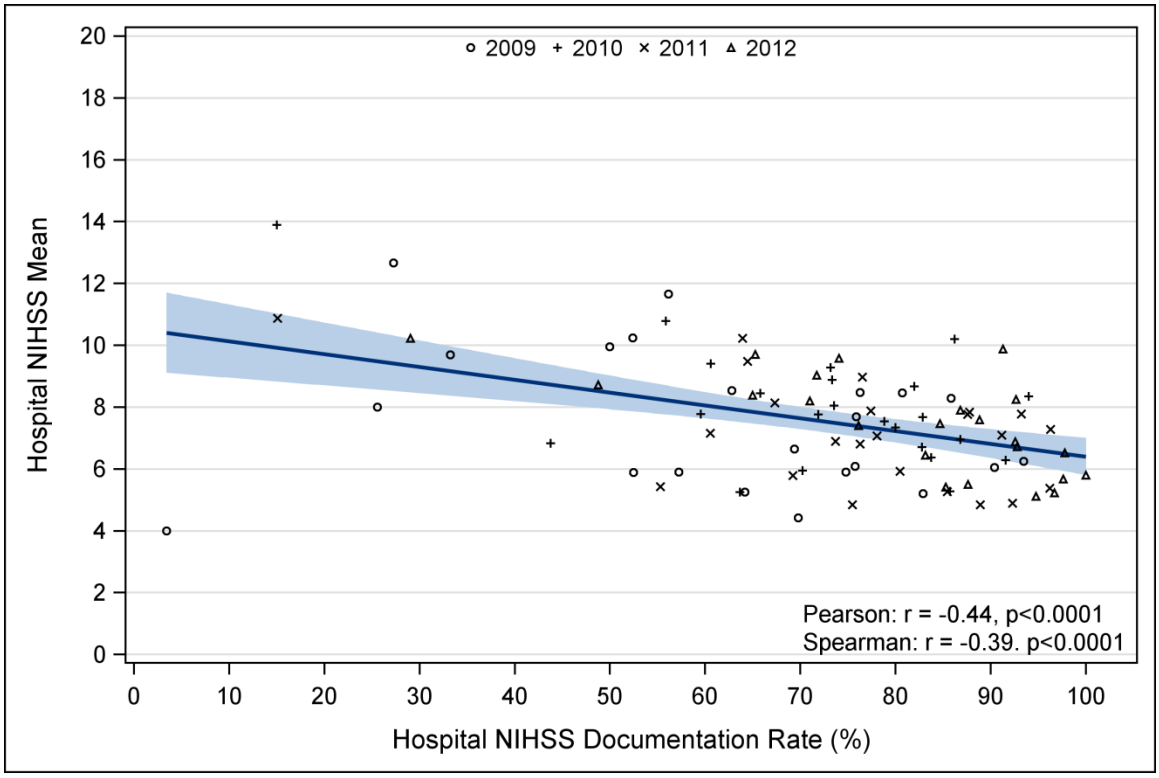

Figure 2.2. Scatter plot of aggregated mean hospital NIHSS score vs. hospital NIHSS documentation rate with fitted regression line (95% CI) in each year (2009-2012).

Figure 2.3 overlays the patient-level distribution of NIHSS scores stratified by the tertile of hospital documentation (<70%, 70-85%, ≥85%). Both ANOVA ($F=14.4$, $df=2$, $p<0.0001$) and Kruskal-Wallis tests (chi-square=64.5, $df=2$, $p<0.0001$) found statistically significant differences in NIHSS score distributions between tertiles of hospital-level documentation rate. The kernel

density curves confirm our findings, with the lower levels of hospital-level NIHSS documentation resulting in slightly higher reported NIHSS scores (i.e. a "shift to the right"). Both of these findings suggest that missing NIHSS data may be MNAR.

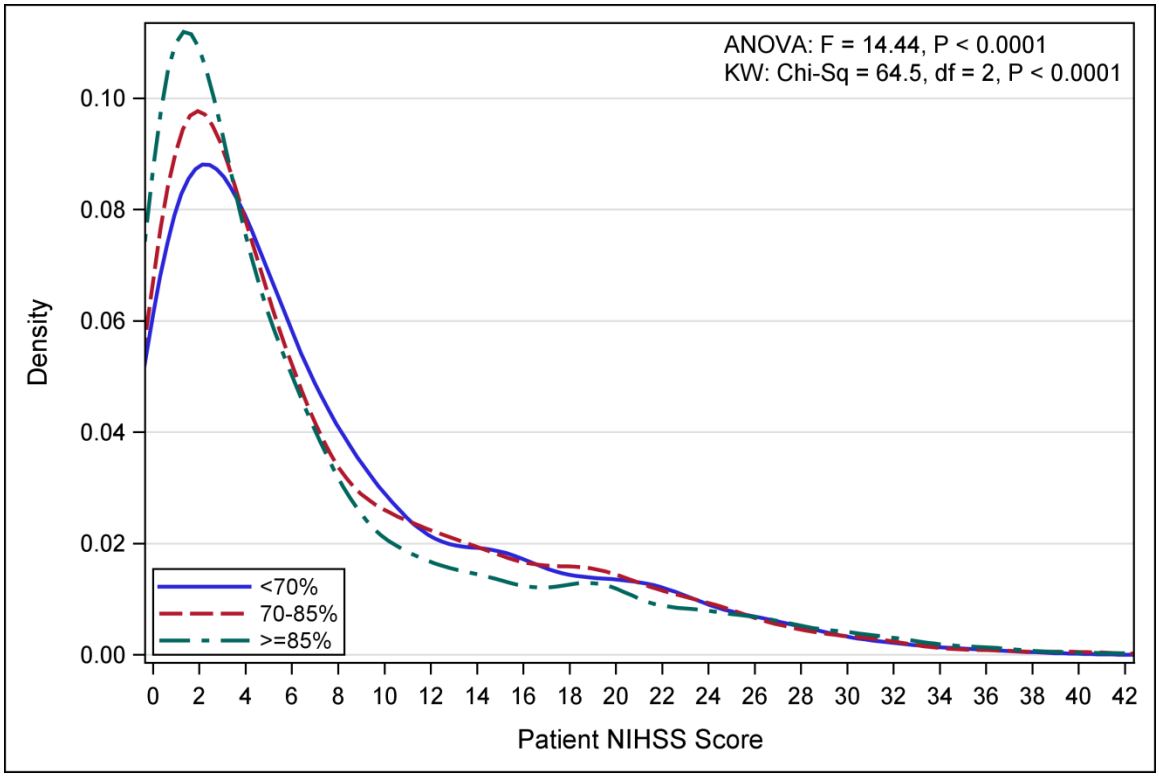

Figure 2.3. Kernel density curves for patient distribution of NIHSS score, stratified by tertile of hospital NIHSS documentation rate (<70%, 70-85%, ≥85%), with ANOVA and Kruskal-Wallis (KW) test results.

Aim 1 – Discussion

The purpose of this study was to investigate patient- and hospital-level patterns and predictors of NIHSS documentation. Our study confirmed our hypothesis that patients with documented NIHSS are not simply a random sample of all stroke patients, and suggests that missing NIHSS data may be MNAR. Our data also suggest that documentation of NIHSS is a

reflection of both patient-level factors, including stroke severity, and the overall hospital-level documentation at the facility in which a patient is treated.

From the hierarchical logistic regression model, we found that patients whose symptoms had resolved by arrival had roughly one-tenth the odds of documentation compared to patients still experiencing stroke symptoms. If it is assumed that resolution of symptoms is accurately recorded at ER arrival, then such patients could be imputed to NIHSS=0. In patients with observed NIHSS, the median (IQR) NIHSS for patients whose symptoms had resolved was as expected 0 (0-2). However, 32% of patients had an NIHSS greater than 0. Thus, imputation may be a feasible solution in improving the overall documentation of NIHSS, given that 16% of undocumented cases had symptoms resolve by arrival.

In a previous study, we found that documentation of NIHSS reflected patients who were candidates for tPA.[66] In this study, we also found that tPA administration was much higher in patients with documented NIHSS than those with undocumented NIHSS (9.3% vs. 1.0%). This may also explain why patient who were received in the ER had greater odds of documentation, as they are typically initial candidates to receive tPA. Similarly, patients who were transferred had higher odds of documentation compared to patients who arrived by EMS or private transportation, as they most likely represent more severe patients. Any effect of EMS was most likely accounted for by the variable for arrival to the ER. Additionally, we found that patients with atrial fibrillation and dyslipidemia had higher rates of documentation. These factors may also be proxies for more severe strokes, as atrial fibrillation[90] and dyslipidemia[91] are significant predictors of stroke severity.

No hospital-level characteristics significantly predicted documentation, although it appears that documentation is higher in hospitals with primary stroke center certification from the Joint Commission. Low power due to a small number of hospitals (n=23) with little between-hospital variability in hospital characteristics may explain this finding. We also found a large statistically significant proportion of the variation ($\sigma_j = 1.09, p < 0.0001$; ICC=25%) in NIHSS documentation in our model can be attributed to the hospital level, suggesting that patient-level NIHSS documentation is also a reflection of overall hospital-level NIHSS documentation.

Analyses of hospital-level documentation and NIHSS scores also confirmed our hypothesis. At the hospital-level, increased documentation of NIHSS was moderately correlated with lower mean NIHSS scores (Pearson correlation r = -0.44; Spearman correlation, r = -0.39). This suggests that hospitals with lower documentation of NIHSS may be underreporting less severe strokes. This was also reflected in our kernel density curves, which showed a "right shift" in the patient distribution of NIHSS in lower documenting hospitals. Our data also shows that hospital-level NIHSS documentation has greatly improved from 2009 to 2012, which is a promising trend if NIHSS is to be used in risk adjustment models.

If missing data were not associated with any characteristics, i.e., a truly random sample of patients, it would be considered missing completely at random (MCAR), and estimates from a complete case analysis are less subject to bias.[67,68] Since we identified significant predictors of NIHSS documentation, we can eliminate the possibility that missing NIHSS data is MCAR. The other possibility is that data is missing at random (MAR), which is to say it is related to some observed variable, but not the value of the missing data itself.[67,68] In cases of either MAR and

MNAR, estimates from a complete case analysis may be biased, however, statistical methods such as multiple imputation or maximum likelihood estimation are often used to correct bias from data that is MAR.[68,92-95] It should be noted that no statistical methods can distinguish between MAR and MNAR mechanisms of missing data. Furthermore, it is possible that missing NIHSS data may be both a combination of MAR and MNAR mechanisms. However, given that our hierarchical model suggests that characteristics of severe stroke patients are associated with documentation, and that increased hospital-level documentation is associated with a shift towards less severe patients, we suspect that NIHSS may be MNAR.

Accurate risk adjustment in hospital profiling requires that variables used in risk adjustment model be of sufficiently high quality.[22] There is already concern that the quality of NIHSS in risk adjustment models due to poor documentation may not be adequate.[26] With the recent announcement that NIHSS is to be included in ICD-10 coding, there will be substantial pressure to include stroke severity in risk adjustment models using administrative data. Based on our evidence, it should be recognized that any hospital-level performance measure that includes NIHSS in risk adjustment is potentially biased if missing NIHSS data are present. Further research is needed to assess the extent of bias in hospital-level mortality measures when cases with undocumented NIHSS are excluded from risk adjustment models and profiling, especially if NIHSS data are MNAR.

There are limitations in this study that should be considered. The sample of hospitals used in this study is a subsample of all Michigan hospitals, which may not be representative of all Michigan hospitals or hospitals nationwide. A greater proportion of MSR patients go to teaching hospitals (93% vs. 61%) and Joint Commission primary stroke center hospitals (78% vs.

65%) compared to patients in the GWTG-Stroke nationwide registry.[96] Thus, patients in the MSR may be more similar to each other compared to what may be seen in the GWTG-Stroke registry or administrative claims data. Replication of this analysis in a larger sample of hospitals may provide more generalizable results, and provide better estimates about the hospital characteristics related to NIHSS documentation. Furthermore, it would be advantageous to repeat this analysis in the future, given the improving documentation of NIHSS over time. Finally, this analysis was conducted in a stroke registry setting, which has clearly defined data abstraction procedures. Further research should be done to assess the completeness and validity of NIHSS in administrative data.

In summary, despite recent improvements in documentation of NIHSS, our evidence suggests that patients with documented NIHSS are a biased subsample of all ischemic stroke patients. Documentation of NIHSS is associated with more severe stroke patients, and is also a reflection of overall hospital-level documentation of NIHSS. Given that NIHSS is a strong predictor of patient outcomes, further study should be done to assess the degree of bias in hospital profiling when a subsample of patients is used to calculate hospital performance measures. Unless complete documentation of NIHSS is achieved, this limitation should be considered when using NIHSS in risk adjustment models.

CHAPTER 3: ASSESSING SELECTION BIAS IN PATIENTS WITH DOCUMENTED NIHSS USING THE HECKMAN SELECTION MODEL

Aim 2 - Background

The missing data problem is common in clinical research.[68,70] Excluding observations with missing data in statistical models, i.e. performing a complete case analysis, has been shown to bias model estimates.[72-74] Missing data are especially pervasive in administrative datasets such as billing data or electronic health records, where variables are frequently undocumented.[71]

Measures of stroke severity, such as the National Institutes of Health Stroke Scale (NIHSS), are strong predictors of patient outcomes.[54,56,84] Currently, it is collected solely in clinical registries where it is frequently underreported, and is absent from administrative data.[54] However, it was recently announced that NIHSS is to be included in ICD-10 coding, with the intent to include NIHSS in stroke performance measures using administrative data. But, if NIHSS is underreported in administrative data, then excluding patients without NIHSS documented may introduce bias into models of hospital performance, if it is a biased subsample of patients, i.e. NIHSS data are missing not at random.[67,68]

Using the Heckman Selection Model, we will test for the presence of selection bias in patients with documented NIHSS. The Heckman Selection Model (hereinafter referred to as the Heckman model), was pioneered by James J. Heckman in 1979 to identify and correct for bias in study estimates resulting from a non-randomly selected sample.[97] He illustrated that when estimating wages of women in the workforce, the population of women excluded housewives, who had self-selected out of the workforce. Thus, the distribution of wages was truncated

because it excluded a group of women for whom wages were not sufficient for them to enter the workforce. Previously, other methods – such as identifying patterns and predictors of documentation – were used to provide evidence of selection bias, but ultimately, investigator intuition was used to identifying potential selection bias. The Heckman model offered a method to estimate the magnitude of selection bias in the sample, and importantly, could then be used to adjust outcomes for the potential bias. While the Heckman model is commonly used in economics and social sciences, it has been used sparingly in the biomedical sciences or health services research. Examples of its use to assess and control for survey nonresponse bias include assessments of medication use[98], estimates of HIV prevalence[99], and self-reported quality of life.[100]

The Heckman model consists of a two-equation model with a model predicting the outcome of interest – the outcome model – and a model predicting whether the outcome was observed or not – the selection model. The outcome model is a linear regression model with a normally distributed, continuous dependent variable, and set of independent predictors (x_i). The selection model is a probit model with binary dependent indicator ($R_i = 1$ if the outcome is observed, $R_i = 0$ if unobserved) and set of independent predictors, which typically include the predictors from the outcome model (x_i), as well as additional predictors of NIHSS documentation (ω_i). As opposed to the logistic model typically used for binary outcomes, the probit model is necessary because the Heckman model requires the two equations have jointly normally distributed error terms. The overall model can be seen below.

$$Outcome\ Model\colon\ NIHSS^{*} = x_i\beta + \varepsilon_i,$$

$$where\ NIHSS*\ is\ the\ true\ score\ and\ NIHSS = NIHSS*\ when\ observed.$$

$$Selection\ Model:\ R_i^* = x_i\beta + \omega_i\gamma + u_i,\ \ R_i \begin{cases} = 1\ if\ R_i^* > 0 \\ = 0\ if\ R_i^* \leq 0 \end{cases}$$

Both selection and outcome models have error terms with mean zero: $\varepsilon_i \sim N(0, \sigma^2)$ and $u_i \sim N(0,1)$. When the available data is a random sample of all data, i.e. no selection bias, the error terms are assumed to be independent, and the correlation between error terms is thus $\rho = corr(\mu_i, \varepsilon_i) = 0$. However, in the presence of selection bias, the available data is determined by a sample selection process, which means that the outcome model is dependent on the selection process, which is reflected through correlation between the error terms, i.e., $\rho = corr(\mu_i, \varepsilon_i) \neq 0$.

Typically, the next step in the Heckman model is to obtain a correction factor termed the inverse Mills Ratio using the error term correlation, and include the correction factor in the outcome model to adjust for selection bias. However, previous work has shown that using the correction factor may worsen, rather than improve estimates, especially when significant selection bias has been found or if the selection model is incorrectly specified.[101,102] For the purposes of this analysis, we are simply interested in the Heckman model's utility as a diagnostic test for selection bias, rather than to use this estimate to adjust for selection bias. It is suggested that an *a priori* understanding of possible drivers of selection bias, and the direction and magnitude in which selection bias may occur improves the validity of the method.[101,103]

In Chapter 2, we illustrated that NIHSS documentation may be greater in more severe stroke cases as compared to less severe strokes, but the differences observed were modest. The purpose of this aim is to provide further evidence that NIHSS data are MNAR i.e., that NIHSS documentation is correlated with the NIHSS score by using the Heckman Selection

Model. We hypothesize that a significant correlation coefficient between the outcome model (patient-level NIHSS score) and selection model (patient-level NIHSS documentation) will be detected. Figure 3.1 illustrates the conceptual framework of the Heckman model for this analysis.

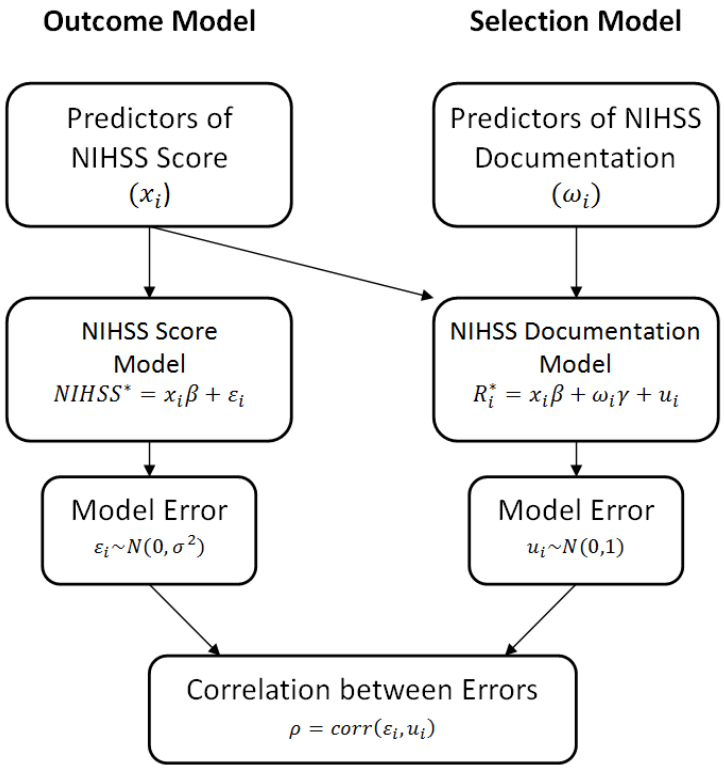

Figure 3.1. Conceptual Framework of Heckman Selection Model in this analysis.

Aim 2 – Methods

Data and Participants

We will again use data from the Michigan Stroke Registry (MSR) from 2009-2012 as described in Chapter 2. Briefly, the MSR is a statewide clinical registry which originated as a prototype for the Paul Coverdell National Acute Stroke Registry, and has been described elsewhere.[85] Currently, the MSR is used to provide a data driven approach to improve the quality of stroke care in the State of Michigan.[86,87] The MSR collects information on many

different patient level characteristics including demographics, EMS and hospital admission information, and clinical information such as stroke severity, ambulatory status, and medical history. In addition, we obtained hospital characteristics from the American Hospital Association annual survey[88] and Paul Coverdell National Acute Stroke Registry hospital inventory.

We used MSR data from 2009 to 2012 for this analysis. To increase the comparability of our findings to a CMS ischemic stroke population, we applied a number of exclusions to the MSR data. Ischemic stroke patients were included if they were aged 65 years or older and excluded cases if they belonged to a hospital with <25 annual cases of ischemic stroke, which is the minimum number of cases for a hospital risk-standardized mortality rate (RSMR) to be calculated, as defined by CMS.[18] We also excluded patients if the stroke occurred in a hospital inpatient setting. As this study was a secondary analysis of deidentified registry data, it was considered exempt from Institutional Review Board review. All analyses were conducted with the use of SAS version 9.3 (SAS Institute Inc, Cary, NC).

Predictor Variables

We examined a number of patient-level predictors in the outcome and selection models. Demographic characteristics included: age, gender (male vs. female), race (white, black, other, not documented), and insurance status (Medicare, Medicaid, private, none). We also assessed EMS and hospital admission information, such as: place stroke occurred (at home vs. in a healthcare setting), arrival mode (EMS, private, transferred), arrival into the ER (yes vs. no), symptoms resolved prior to arrival (yes vs. no), and tPA administration (yes vs. no). Finally, we also examined several clinical variables in this analysis, including: able to ambulate pre-

stroke, diabetes mellitus, congestive heart failure, peripheral artery disease, hypertension, current smoker, and history of prior stroke, transient ischemic attack/vertebrobasilar insufficiency (TIA/VBI), or myocardial infarction/coronary artery disease (MI/CAD).

Hospital characteristics were also examined and included annual stroke volume (<200, 200-600, 600+), urban vs. rural location, teaching status, presence of an acute stroke team, and Joint Commission primary stroke center status.[63]

Outcome Model Specification

The dependent variable in the outcome model is the patient-level NIHSS score, which is continuously distributed with a 0-42 point integer scale. A primary assumption of the outcome model in the Heckman model is that the dependent variable be a normally distributed, continuous variable.[97] To satisfy this assumption, we transformed the NIHSS score to a normal distribution using a Box-Cox transformation, as the NIHSS distribution is highly right-skewed. The outcome model was specified using a backward selection process of predictor variables with stepwise deletion of non-significant predictors. Quasi-maximum likelihood estimation was used to produce robust estimates of standard errors to account for a clustering effect of patients within hospitals.

Selection Model Specification

The dependent variable in the selection model is the patient-level binary indicator of NIHSS documentation (documented vs. undocumented). The selection model was specified to include all significant predictors of NIHSS score, regardless of statistical significance in the selection model. Additional significant predictors of NIHSS documentation were again derived from a backward selection process of the remaining predictor variables with stepwise deletion

of non-significant predictors. Again, we calculated robust standard errors using quasi-maximum likelihood estimation to account for clustering of patients in hospitals.

Estimating the Correlation Coefficient

Once the outcome and selection models were specified, we utilized the PROC QLIM procedure in SAS to estimate the correlation coefficient between the outcome model and selection model error terms. The PROC QLIM (qualitative and limited dependent variable model) procedure allows users to estimate the correlation between a simultaneously specified multivariable outcome and selection model.[104] A statistically significant correlation coefficient would indicate the presence of selection bias in patients with documented NIHSS. The correlation coefficient ranges from +1 to -1, with 0 representing no selection bias, and +/-1 representing strong selection bias. A positive correlation would indicate that as the NIHSS score increases, i.e. strokes are more severe, documentation increases. Conversely, a negative correlation indicates that as NIHSS score increases, documentation decreases.

To investigate how the prevalence of missing NIHSS data impacts the amount of selection bias in the sample, we repeated the analysis using only data from 2009-2010, when documentation was lower (68%), and again using only data from 2011-2012, when documentation was greater (80%), and estimated the correlation coefficient between the models in each time period (2009-2010 and 2011-2012).

Aim 2 - Results

From the Michigan Stroke Registry, we used data from 10,717 ischemic stroke cases discharged from 23 hospitals for the analysis, of which 7,957 cases (74.3%) had NIHSS documented. The following variables were statistically significant independent predictors of

NIHSS score in the outcome model: age (in years), gender (male vs. female), stroke occurred at home vs. in a healthcare setting, symptoms resolved prior to arrival, mode of arrival (EMS, private, transfer), tPA administration (yes vs. no), ambulatory pre-stroke (yes vs. no), presence of an acute stroke team (yes vs. no), and history of atrial fibrillation, prior stroke, dyslipidemia, and heart failure (yes vs. no). (Table 3.1) A positive beta coefficient indicates that the variable is associated with a more severe stroke, while a negative beta coefficient is associated with a less severe stroke. When these variables were included in the selection model, all were statistically significant predictors of documentation except age, gender, and PMH heart failure. The selection model included the following additional significant predictors of NIHSS documentation: insurance status (Medicare, Medicaid, Private, none), race (white, black, other, not documented), patient received in the ER (yes vs. no), year (2009, 2010, 2011, and 2012), hospital Joint Commission primary stroke center status (yes vs. no), and hospital stroke volume (<200, 200-600, 600+). (Table 3.1)

Table 3.1. Heckman Selection Model specifications for outcome and selection models in the full sample of n=10,717 stroke cases (2009-2012).

Variable	***Frequency***	***Outcome Model***		***Selection Model***	
	% or Mean(SD)	*Coefficient (SE)*	*p-value*	*Coefficient (SE)*	*p-value*
Intercept	-	1.4828 (0.1261)	<0.0001	1.3778 (0.2089)	<0.0001
Age (per year)	78.7 (8.3)	0.0103 (0.0013)	<0.0001	-0.0022 (0.0015)	0.1547
Gender (Male)	54.7	-0.0625 (0.0202)	0.0020	-0.0308 (0.0240)	0.1992
Stroke Occurred at Home	89.5	-0.1612 (0.0357)	<0.0001	0.1842 (0.0379)	<0.0001
Symptoms Resolved	6.6	-0.8008 (0.0501)	<0.0001	-0.9348 (0.0310)	<0.0001
Arrival Mode					
EMS	49.5	-0.0239 (0.0283)	0.3986	-0.1055 (0.0412)	0.0105
Private	34.9	-0.6342 (0.0293)	<0.0001	-0.1030 (0.0416)	0.0134
Transfer	42.6	Ref	-	Ref	-
tPA Administered	7.2	0.8149 (0.0393)	<0.0001	1.1631 (0.0885)	<0.0001
Ambulatory Pre-stroke	95.1	-0.4972 (0.0536)	<0.0001	0.3147 (0.0517)	<0.0001
History of Atrial Fibrillation	22.7	0.2319 (0.0244)	<0.0001	0.0943 (0.0303)	0.0018
History of Prior Stroke	27.7	0.1515 (0.0226)	<0.0001	-0.0604 (0.0258)	0.0193

Table 3.1. (cont'd) Heckman Selection Model specifications for outcome and selection models in the full sample of n=10,717 stroke cases (2009-2012).

Variable	***Frequency***	***Outcome Model***		***Selection Model***	
	% or Mean(SD)	*Coefficient (SE)*	*p-value*	*Coefficient (SE)*	*p-value*
History of Dyslipidemia	45.5	-0.1005 (0.0203)	<0.0001	0.1263 (0.0239)	<0.0001
History of Heart Failure	13.5	0.1609 (0.0299)	<0.0001	-0.0327 (0.0245)	0.3433
Acute Stroke Team	85.3	0.1076 (0.0276)	<0.0001	-0.1458 (0.0362)	<0.0001
Insurance Status					
Medicare	49.1	-	-	0.2944 (0.1220)	0.0158
Medicaid	5.2	-	-	0.1043 (0.1299)	0.4222
Private	45.0	-	-	0.2496 (0.1225)	0.0417
None	0.7	-	-	Ref	-
Race					
White	72.6	-	-	-0.0758 (0.0456)	0.0964
Black	18.1	-	-	-0.3237 (0.0503)	<0.0001
Other	1.1	-	-	0.0627 (0.1285)	0.6258
Not Documented	8.2	-	-	Ref	-
Received in the ER	88.3	-	-	0.2424 (0.0431)	<0.0001
Year					
2009	25.0	-	-	-0.5756 (0.0346)	<0.0001
2010	23.5	-	-	-0.3620 (0.0458)	<0.0001
2011	26.0	-	-	-0.2103 (0.0359)	<0.0001
2012	25.5	-	-	Ref	-
PSC Status	77.9	-	-	0.2211 (0.0294)	<0.0001
Stroke Volume					
<200	9.0	-	-	0.2675 (0.0443)	<0.0001
200-600	38.6	-	-	0.3344 (0.0283)	<0.0001
600+	52.3	-	-	Ref	-

Table 3.2 shows the estimated correlation coefficients between the error terms of specified outcome and selection models. For the entire sample, we estimated a statistically significant correlation coefficient of ρ=0.11 (95% CI: 0.09, 0.13; p <0.0001). (Table 3.2) This is interpreted as weak, but statistically significant, selection bias. The positive sign on the correlation indicates that as NIHSS score increases, the probability of documentation also increases. When we restricted data to 2009-2010, when documentation was relatively lower (68%), we found a slightly higher correlation coefficient of ρ=0.13 (95% CI: 0.07, 0.20; p<0.0001), indicating a modest increase in selection bias when documentation was lower.

Conversely, when we limited our data to 2011-2012, when documentation was better (80%), we found a slightly lower correlation coefficient of ρ=0.07 (95% CI: 0.05, 0.09; p<0.0001), indicating less selection bias when documentation had improved.

Table 3.2. Estimated correlation coefficient between error terms of outcome and selection models for the full sample (2009-2012), and by 2009-2010 and 2011-2012.

Sample	*Total*	*Num. NIHSS Documented, n (%)*	*Estimated Correlation Coefficient (95% CI)*	*p-value*
2009-2012	10,717	7,957 (74.3)	0.11 (0.09, 0.13) [a]	<0.0001
2009-2010	5,197	3,554 (68.4)	0.13 (0.07, 0.20) [b]	<0.0001
2011-2012	5,520	4,403 (79.8)	0.07 (0.05, 0.09) [c]	<0.0001

[a] Outcome and selection model variables can be seen in Table 3.1.

[b] Outcome model variables: age, gender, stroke occurred at home, arrival model, received in the ER, symptoms resolved by ER arrival, tPA administration, ambulatory pre-stroke, and history of atrial fibrillation, heart failure, stroke and myocardial infarction.
Selection model variables: Outcome model variables plus race, history of smoking, year, and hospital stroke volume, rural location, Joint Commission Primary Stroke Center status, presence of acute stroke team.

[c] Outcome model variables: age, stroke occurred at home, arrival mode, symptoms resolved by ER arrival, tPA administration, ambulatory pre-stroke, hospital presence of acute stroke team, and history of atrial fibrillation, heart failure, stroke, and TIA/vertebrobasilar insufficiency.
Selection model variables: Outcome model variables plus race, received in the ER, year, and hospital stroke volume, rural location, teaching status, Joint Commission Primary Stroke Center status.

Aim 2 – Discussion

Using the Heckman model, we found evidence of selection bias in patients with documented NIHSS. Although statistically significant, the magnitude of the selection bias appears to be relatively weak (ρ=0.11). The positive correlation between stroke severity (NIHSS score) and NIHSS documentation suggests that as stroke severity increases, the probability of documentation also increases. In Aim 1, we concluded that more severe strokes are better documented. Our results in this study confirm our findings, which also indicate that as patient-level stroke severity increases, the probability of documentation also increases. Furthermore, as expected we found that when documentation of NIHSS was lower, the magnitude of selection bias slightly increased (ρ=0.13), and bias subsequently decreased slightly when

documentation improved (ρ=0.07). In the traditional Heckman model process, we would subsequently calculate the inverse Mills ratio using the estimated correlation coefficient, and include this parameter in the model predicting patient-level NIHSS score to correct for the selection bias. However, since our aim is only to assess for the presence of bias, and not estimate patient-level NIHSS score, we did not perform this step.

Together, this evidence provides a compelling argument that documentation of NIHSS is associated with the NIHSS score itself, and that the probability of documentation increases as patient-level NIHSS increases. As such, it can be safely concluded that missing NIHSS data are missing not at random (MNAR). Therefore, any analysis that includes NIHSS when it is not completely documented is subject to selection bias; however, the extent of the bias appears to be modes and its implications have yet to be understood. Both analyses we conducted (Chapters 2 and 3) suggested that NIHSS data are MNAR; but selection bias appears to be relatively weak. In Aim 1, we showed a slight "right shift" in the distribution of patient-level NIHSS in low documenting hospitals. The average NIHSS score in high versus low documenting hospitals was 6.7 compared to 8.7, respectively. (Table 2.2) But, would using a more severe subsample of stroke patients translate to biased estimates of hospital-level mortality? We will attempt to answer that question in the next aim.

This particular analysis does have some limitations to consider. First, the Heckman model relies on an accurately specified selection model.[103] Failure to specify a correct selection method may result in inaccurate assessment or correction for selection bias. Having *a priori* information about the possible direction of selection bias or what variables might predict selection may improve the validity of Heckman model estimates. The analysis in Aim 1

corroborates our findings, which also suggest a similar magnitude and direction of selection bias, improving the validity of our results. The Heckman model also requires that the dependent variable for the outcome model is a continuous, normally distributed variable.[97] The NIHSS score is not a normally distributed variable (as illustrated in Figure 2.3). However, we used a Box-Cox transformation to transform NIHSS into an approximately normal distribution. As this study used a relatively small subset of hospitals (n=23), further research should be done to improve the generalizability our findings in a more representative sample of hospitals.

Using the Heckman Selection Model, we were able to corroborate previous analyses which showed evidence of weak selection bias in patients with NIHSS documented. We also illustrated that as documentation of NIHSS improved the magnitude of selection bias in our data reduced. It is unclear if hospital-level performance measures (e.g. mortality) are biased when documentation patterns change in respect to patient stroke severity. In the next chapter, we will employ computer simulations to explore how the prevalence and mechanism of missing NIHSS data impacts the accuracy of hospital performance profiling.

CHAPTER 4: THE IMPACT OF MISSING NIHSS DATA ON THE ACCURACY OF HOSPITAL PROFILING

Aim 3 – Background

While it is widely accepted that using a complete case analysis in the presence of missing data may introduce bias into any given analyses[68,69], how this impacts hospital-level estimates used for hospital profiling is less certain. There is some evidence that hospital-level measures of performance may be biased when missing data is present. One simulation study comparing risk-adjusted hospital trauma-related mortality measures showed that a complete case analysis when risk-adjustment variables were missing not at random (MNAR) led to considerable changes in hospital-level mortality profiling.[64] Another simulation study examining the impact of missing data on profiling of pay-for-performance outcomes showed that between 11 to 21 percent of misclassification was attributable to missing data used in risk adjustment models.[65] Studies have also shown that differential coding[40] or undercoding[82] of comorbidities and severity indicators between hospitals – which would cause "missing" data if variables were coded incorrectly – can bias hospital-level risk standardized mortality rates.

The addition of 30-day mortality and readmission measures for ischemic stroke into the Centers for Medicare & Medicaid Services pay-for-performance schemes[5,6] has generated considerable contention regarding the contents of models used in risk adjustment. There is serious concern that excluding a measure of stroke severity, such as the National Institutes of Health Stroke Scale (NIHSS)[50,51], will not adequately risk adjust hospital-level performance measures.[53-56] Furthermore, it has been suggested that hospitals which tend to see a more severe case-mix of patients – such as tertiary referral centers or primary stroke centers – may

be at greater risk of misclassification.[46-48] With the announcement that NIHSS is to be included in ICD-10 coding, it will likely be included in future risk adjustment models for ischemic stroke outcomes. Although documentation of NIHSS in clinical datasets has improved in recent years, it is still frequently missing in large clinical datasets.[53] Therefore, it is essential to understand how missing NIHSS data may impact hospital-level estimates of mortality used in profiling schemes, especially if it is MNAR.

In this study, we utilize computer simulations to illustrate how missing NIHSS data impacts the accuracy of hospital performance profiling on ischemic stroke mortality. Specifically, we will assess how the prevalence and mechanism by which NIHSS is missing impacts our ability to classify hospital outliers, estimate hospital deviation in ischemic stroke mortality, and correctly rank-order hospitals on ischemic stroke mortality. We hypothesize that our ability to correctly identify outlier hospitals and rank-order hospitals will degrade as the prevalence of missing NIHSS increases, especially in situations where missingness is related to the severity of the stroke, i.e. is MNAR. Finally, because hospital case volume has previously been shown to impact profiling accuracy in myocardial infarction, we will also investigate how hospital ischemic stroke volume modifies our findings.

Aim 3 – Methods

To pursue our aims, data must be generated in such a way that the variation in patient case-mix and ischemic stroke mortality within and between hospitals reflect empirical estimates from real-world data. Briefly, a top down approach for data generation was used, where a set of hospitals were generated with assigned components of case-mix and ischemic stroke mortality variation. Patients were then generated within each hospital, and assigned

patient characteristics that reflect the underlying observed case-mix and mortality. We then replicated the generated dataset, simulated missing NIHSS data within each dataset based on a different mechanism and prevalence of missing NIHSS data. Hospital-level outlier status and RSMRs were estimated from a complete case analysis of patients with observed NIHSS. This data generation scheme can be seen in Figure 4.1

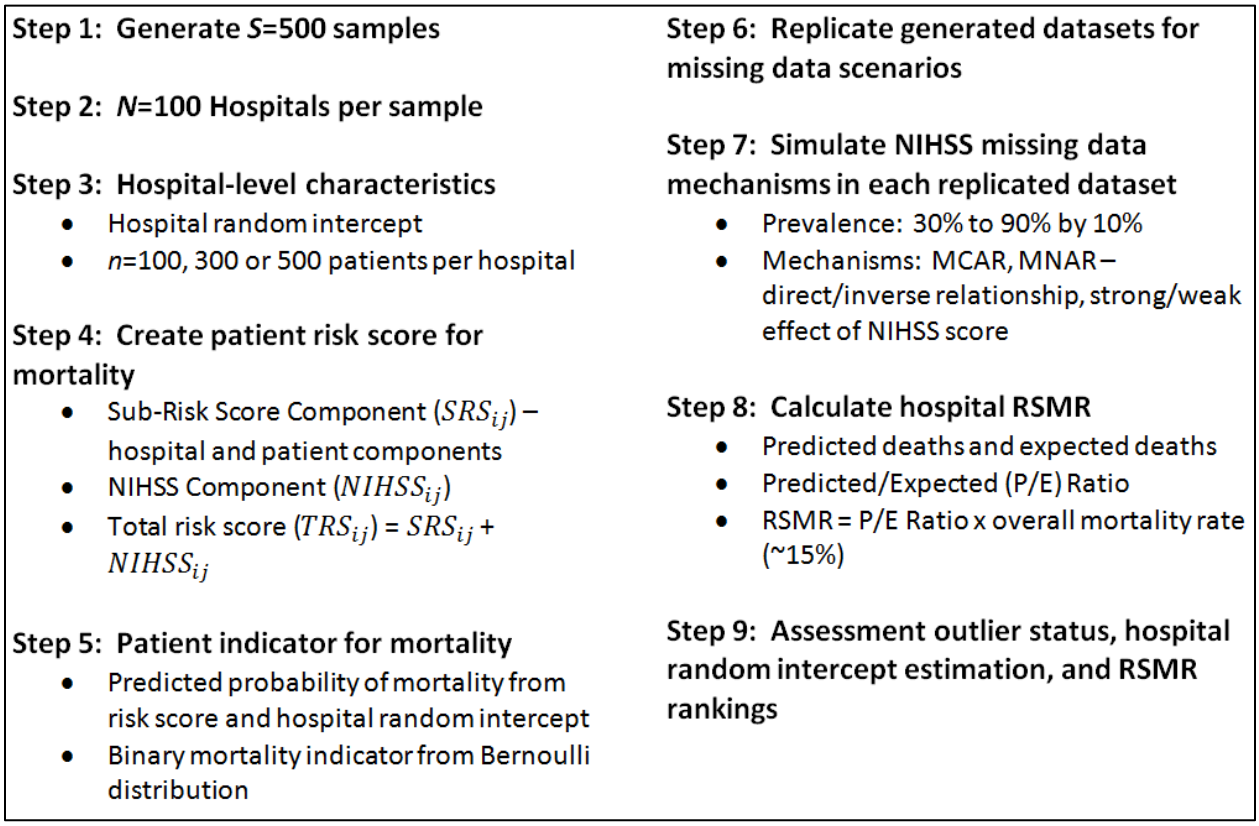

Figure 4.1. Overview of data generation process for simulations.

Section 1 - Parameter Generation for Simulations

A series of analyses of 10,717 ischemic stroke patients 65 years of age and older from 23 hospitals in the Michigan Stroke Registry (MSR) were conducted to generate parameters needed for the computer simulations. The MSR is a statewide clinical registry which originated

as a prototype for the Paul Coverdell National Acute Stroke Registry, and has been described elsewhere.[85] Descriptive statistics of the sample can be seen in Table 2.1.

Parameters needed for the simulation studies were generated in three distinct steps: first we created a multivariable patient risk score for in-hospital mortality using MSR data. Second, we quantified the variation in the patient risk score between hospitals in the registry (this variation represents the differences in hospital case-mix). Finally, we estimated hierarchical model parameters for in-hospital mortality model given the patient risk score and hospital random intercepts. Specific details of the steps are described below.

Patient Risk Score: We used the Get With the Guidelines – Stroke (GWTG-Stroke) in-hospital mortality risk score for this analysis, which includes NIHSS score.[54] In-hospital mortality was used because the MSR does not have data on 30-day mortality. However, for acute myocardial infarction patients in-hospital mortality has been shown correlate well with 30-day mortality.[105] The GWTG-Stroke in-hospital mortality risk score was developed from the logistic model using the method described by Sullivan, et al.[106], and contains nine clinical variables: patient age, NIHSS score categories (0-2, 3-5, 6-10, 11-15, 16-20, 21-25, and 26-42), mode of arrival, gender, and presence of atrial fibrillation, previous stroke or TIA, coronary artery disease, diabetes mellitus, or history of dyslipidemia. (Table 4.1) The score ranges from 0 to 109, and is shown in Table 4.1; the NIHSS score is by far that most important variable contributing to the total score

Table 4.1. Get With the Guidelines-Stroke in-hospital mortality risk score variables, categories, and respective points.

Variable	***Categories***	***Points***	
Age (in years)	<60	0	
	60-70	1	
	70-80	5	
	≥80	9	
NIHSS Score	0-2	0	
	3-5	10	
	6-10	21	
	11-15	37	
	16-20	48	
	21-25	56	
	26-42	65	
Mode of Arrival	Private transport	0	
	Did not present via ED	16	
	Ambulance from scene	12	
		Yes	**No**
Presence of:	Male gender	0	3
	Atrial fibrillation	5	0
	Previous stroke or TIA	0	2
	Coronary artery disease	5	0
	Diabetes mellitus	2	0
	History of dyslipidemia	0	2

Information taken from Smith, et al. 2010[54]

To allow for manipulation of the NIHSS score variable, we calculated the NIHSS risk score component separately from the rest of the risk score. Therefore, the total risk score (TRS_{ij}) for patient *i* in hospital *j* is the sum of the NIHSS score component ($NIHSS_{ij}$) and a non-NIHSS component – hereinafter referred to as the sub-risk score component (SRS_{ij}) – which contains the remaining eight variables. (1)

$$(1)\ TRS_{ij} = SRS_{ij} + NIHSS_{ij}$$

In the MSR, the sub-risk score (SRS_{ij}) is normally distributed with mean 21.4 and standard deviation (SD) 8.3, i.e. $SRS_{ij} \sim N(21.4, 8.3^2)$. The distribution of NIHSS score categories in the 7,957 (74.3%) cases with documented NIHSS in the MSR can be seen in Figure 4.2. The

mean (SD) and median (IQR) for patients with documented NIHSS were 7.3 (SD=7.8) and 4 (IQR=2-11), respectively.

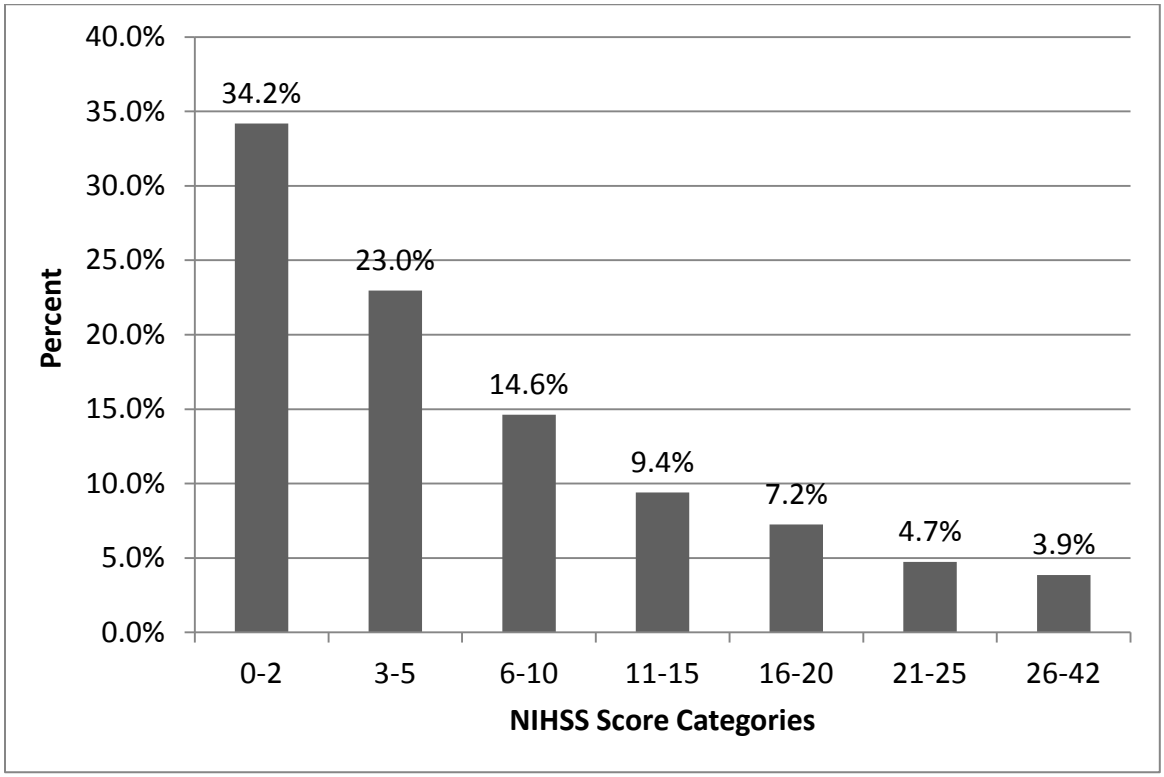

Figure 4.2. Distribution of patient-level NIHSS score categories in the Michigan Stroke Registry (n=7,957)

To measure the association between SRS_{ij} and $NIHSS_{ij}$ components, we used an ordinal regression model to predict NIHSS score categories given the patient SRS_{ij}. For simplicity in the simulation process, we used an ordered probit model, which has a normally distributed random error term, as opposed to the traditional ordinal logistic model, where the error term has a logistic distribution. The ordered probit model yields a beta coefficient for the sub-risk score, and six intercept terms, which reflect the cutoff points between the seven ordinal NIHSS categories. (Table 4.2)

Table 4.2. Results of ordered probit model of NIHSS category predicted by sub-risk score. (n=7,957)

Parameter	*Estimate*	*Standard Error*
Intercept 1*	0.63	0.0353
Intercept 2	1.26	0.0365
Intercept 3	1.79	0.0381
Intercept 4	2.16	0.0396
Intercept 5	2.55	0.0419
Intercept 6	2.98	0.0459
Sub-Risk Score (SRS_{ij})	-0.050	0.00152

Note: All parameter estimates are statistically significant, p<0.0001

* Intercepts reflect cutoff points between seven ordinal NIHSS categories, 0-2, 3-5, 6-10, 11-15, 16-20, 21-25, and 26-42.

Using the model intercepts shown in Table 4.2, patient-level NIHSS category can be imputed by multiplying the generated sub-risk score and sub-risk score model beta coefficient (β=-0.050). This step will be explained in more detail in the section describing the data simulation process.

Between-Hospital Variation in Risk Score: To estimate the between-hospital variation in patient risk score, i.e. case-mix variation, we ran a variance components model to estimate the hospital-level variation in the sub-risk score component (SRS_{ij}) which was centered with mean of 0. The variance of the sub risk score was made up of a hospital-level component, μ_j with variance σ_μ^2, for hospital *j*, and a patient-level component, δ_{ij} with variance σ_δ^2, for patient *i* in hospital *j*, $SRS_{ij} = \mu_j + \delta_{ij}$.

It is assumed that the hospital and patient-level variance components are independent from one another. Thus, $var(SRS_{ij}) = \sigma_\mu^2 + \sigma_\delta^2$. From the variance components model, we estimated $\mu_j \sim N(0, \sigma_\mu^2 = 1.5)$ and $\delta_{ij} \sim N(0, \sigma_\delta^2 = 68.0)$. Using the formula for calculating intraclass correlation coefficient – $ICC = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\delta^2) = 1.5/(1.5 + 68.0) = 0.022$, or 2.2%. This means that only 2.2% of the variation in the sub-

risk score is attributed to between-hospital differences in the overall mean sub risk score. Because the NIHSS component is estimated from the sub-risk score component, case-mix variation in NIHSS will also be reflected by variation in the sub-risk score.

Between-Hospital Variation in In-Hospital Mortality: Using data from the MSR, a hierarchical logistic regression model was used to estimate the between-hospital variation in mortality, given the patient total risk score (TRS_{ij}). This model can be seen below (2), where p_{ij} represents the probability of in-hospital mortality for patient *i* in hospital *j*, β_0 represents the overall log-odds of mortality, β_1 represents the log-odds of mortality given a one unit increase in the total risk score for patient *i* in hospital *j* (TRS_{ij}), and b_{0j} represents the random intercept for hospital *j*.

$$(2)\ \ logit(p_{ij}) = \beta_0 + \beta_1 * TRS_{ij} + b_{0j}$$

When this model was run in 7,957 ischemic stroke patients who had NIHSS recorded in the MSR, we obtained the following model estimates (3)

$$(3)\ \ logit(p_{ij}) = -6.1 + 0.054 * TRS_{ij} + b_{0j}$$

We estimated the distribution of the hospital random intercept, $b_{0j} \sim N(0, \sigma^2 = 0.13)$, where 0.13 represents the between-hospital variation in in-hospital mortality. The ICC from a logistic regression model is calculated using the equation[107], $ICC = \sigma^2/(\sigma^2 + \pi^2/3) = 0.13/(0.13 + \pi^2/3) = 0.039$, or 3.9%, which means that only 3.9% of the unexplained variation in in-hospital mortality is attributed to between-hospital differences. These estimated parameters were subsequently used to simulate a full dataset which mimics the between- and within-hospital variation in patient risk score and mortality.

Section 2 – Generating Datasets for Simulations

We simulated *S*=500 independent samples of patients within each hospital from the parameters generated in the previous section. In each sample (*S*), *N*=100 hospitals were simulated using *n* patients per hospital; each scenario reflected a unique combination of the NIHSS documentation rate (%), the mechanism of missing NIHSS data, and hospital stroke volume. To assess how the accuracy of performance profiling is modified by the hospital ischemic stroke volume, independent simulations of hospital volumes of *n*=100, 300, and 500 patients were used, to represent low, moderate, and high volume hospitals. Each hospital was assigned a random intercept for mortality, which represents its true deviation in mortality from the overall average, i.e., true hospital performance. Patient-level risk scores for mortality were then simulated to represent within- and between-hospital variation in patient risk of mortality observed in the MSR. Finally, using the assigned hospital random intercept and patient-level risk score, we simulated a binary mortality outcome (alive/died) for each patient. Specific details for these three steps are detailed below.

Assigned Hospital Random Intercept: From the analyses conducted in the MSR, we observed a normal distribution of hospital random intercepts of $b_{0j} \sim N(0, \sigma^2 = 0.13)$. Simulated hospitals were randomly assigned a random intercept from this normal distribution. The assigned random intercept represents a hospital's known deviation in mortality compared to the overall average, after adjusting for the patients risk of mortality. As such, it represents a hospital's true in-hospital mortality compared to the average hospital, and was used as the gold standard for comparison with estimated hospital performance rankings.

Assigned Patient Risk Score: To generate the total risk score for patient *i* in hospital *j* (TRS_{ij}), we first generated the non-NIHSS component of the risk score, i.e. the sub-risk score (SRS_{ij}). As previously mentioned the sub-risk score has a hospital-level component, μ_j, and patient-level component, δ_{ij}. For each hospital, μ_j was randomly drawn from the distribution, $\mu_j \sim \mathrm{N}(0, \sigma_{\mu}^2 = 1.5)$. Within each simulated hospital *j*, for patient *i*, δ_{ij} was randomly drawn from the distribution $\delta_{ij} \sim \mathrm{N}(0, \sigma_{\delta}^2 = 68.0)$. The hospital and patient components were summed to create the SRS_{ij}, which was then centered on the observed mean from the MSR (mean=21.4), which can be seen in equation (4).

$$(4)\ SRS_{ij} = (\mu_j + \delta_{ij}) + 21.4$$

Next, we assigned each patient an NIHSS score category by multiplying the SRS_{ij} with the beta coefficient for SRS_{ij} from the ordered probit model, and added a random error term, ε, drawn from an $N(0, 1)$ distribution. (5)

$$(5)\ \gamma_{ij} = -0.050 * SRS_{ij} + \varepsilon$$

The estimate γ_{ij} was then compared to the cutoff points from the ordered probit model intercepts, which refers to an imputed NIHSS category, as seen in Table 4.3.

Table 4.3. NIHSS category assignment cutoff intervals derived from the ordered probit model predicting NIHSS category given the patient sub-risk score.

Cutoff Interval	*Assigned NIHSS Category*	*Risk Score Points* ($NIHSS_{ij}$)*
$\gamma \geq -0.63$	0-2	0
$-0.63 > \gamma \geq -1.26$	3-5	10
$-1.26 > \gamma \geq -1.79$	6-10	21
$-1.79 > \gamma \geq -2.16$	11-15	37
$-2.16 > \gamma \geq -2.55$	16-20	48
$-2.55 > \gamma \geq -2.98$	21-25	56
$-2.98 > \gamma$	26-42	65

* Risk Score Points from Smith, et al. 2010[54]
Note: γ is calculated using the patient Sub Risk Score (SRS)

Finally, the NIHSS risk score and sub risk score components were summed to create the total risk score (TRS_{ij}), which ranges from 0 to 109 and mimics the distribution observed in the MSR. For instance, if a patient was assigned an $SRS_{ij} = 20$ and random error term $\varepsilon = 0$, the resulting γ_{ij} is: $\gamma_{ij} = -0.050 * 20 + 0 = -1.0$. Thus, the patient would be assigned to an NIHSS category of 3-5 (i.e. $-0.63 > \gamma_{ij} \geq -1.26$), and 10 points would be added to the sub-risk score for a total risk score of $TRS_{ij} = SRS_{ij} + NIHSS_{ij} = 20 + 10 = 30$.

Generating In-Hospital Mortality: Using the logit model in equation (2), we input the assigned hospital random intercept and for each of the *n* patients in the particular sample we generated the logit of the predicted probability of mortality (p_{ij}) from the TRS_{ij}. To reflect 30-day mortality rates (~15%) used in CMS outcome metrics as opposed to in-hospital mortality rates (~4%), we re-scaled the model intercept to generate an average mortality of 15% ($\beta_0 = -4.4$ vs. –6.1). (6)

$$(6)\ logit(p_{ij}) = -4.4 + 0.054 * TRS_{ij} + b_{0j}$$

The reverse logit of this model estimated the predicted probability of mortality (p_{ij}) for each patient in the sample. From this we generated a patient-level binary mortality status (0 if alive, 1 if dead) using a random draw from the Bernoulli distribution.

From these three steps, we generated patient samples within each simulated hospital which reflect empirical estimates of variation in case-mix and ischemic stroke mortality obtained from the MSR. Each patient has a generated risk score for mortality, including an NIHSS component, and binary mortality indicator. In the next section, we discuss models which were used to simulate missing NIHSS in the fully observed dataset.

Section 3 – Missing NIHSS Data Model Specification

In the Chapters 2 and 3, we provided evidence that NIHSS documentation is related to the patient NIHSS score. To replicate missing NIHSS data in our simulated data, we generated a mixture of different prevalences and mechanisms of NIHSS documentation. First, we simulated a mechanism where NIHSS documentation is completely unrelated to the NIHSS score, i.e., data are missing completely at random (MCAR). To replicate a MCAR model of NIHSS documentation, we generated a binary indicator of documentation by a random draw from a Bernoulli distribution. In addition to a fully observed dataset, we created datasets which modified the probability of documentation between 30 and 90% in increments of 10% in addition to the fully observed dataset.

Next, we simulated a mechanism where NIHSS score category and NIHSS documentation are directly related (as NIHSS score category increases, documentation increases) and inversely related (as NIHSS score category increases, documentation decreases). These mechanisms represent a missing not at random mechanism of missing data (MNAR), where the missingness in NIHSS is related to the value of the NIHSS score itself. Logistic regression models were used to estimate the probability of NIHSS documentation (R_{ij}) given the patient NIHSS category ($NIHSS_{ij}$). (7)

$$(7)\ \ logit\left[p\left(R_{ij} = 1|\ NIHSS_{ij}\right)\right] = \ \beta_0 + \beta_1 * NIHSS_{ij}$$

Because we cannot observe these models directly, we estimated the model intercept (β_0) – which represents the overall documentation rate – and the beta coefficient (β_1) for the NIHSS score – which indicates the estimated increase or decrease in odds of documentation by moving up one NIHSS category (0-2, 3-5, 6-10, 11-15, 16-20, 21-25, and 26-42). The signs of

beta coefficients were manipulated to reflect direct and inverse relationships between NIHSS and NIHSS documentation. Additionally, in each scenario, we altered the values of the beta coefficient to reflect a relatively weaker and stronger effect of NIHSS category on documentation. The weaker effect represents a 10% increase or decrease (Beta = +/-0.095) in odds of documentation as NIHSS category increases; the strong effect represents a 25% increase or decrease (Beta = +/-0.225) in odds of documentation as NIHSS category increases. In total, four MNAR models were created (direct-weak, direct-strong, inverse-weak, inverse-strong). Similar to the MCAR model, we altered the model intercepts to reflect overall documentation rate of 30 to 90% in increments of 10%. All missing NIHSS model specifications (MCAR and MNAR), and their estimated NIHSS documentation rates can be seen in Table 4.4.

Table 4.4. Specification for missing NIHSS models, including model parameters and estimated documentation rates in each category of NIHSS.

Scenario		***Model Coefficients***		***Estimated NIHSS Documentation Rates*** *NIHSS Score Category*						
Effect	*% Doc.*	*Intercept*	*Beta*	*0-2*	*3-5*	*6-10*	*11-15*	*16-20*	*21-25*	*25-42*
Missing Completely at Random										
-	*90*	-	-	90	90	90	90	90	90	90
	80	-	-	80	80	80	80	80	80	80
	70	-	-	70	70	70	70	70	70	70
	60	-	-	60	60	60	60	60	60	60
	50	-	-	50	50	50	50	50	50	50
	40	-	-	40	40	40	40	40	40	40
	30	-	-	30	30	30	30	30	30	30
Missing Not at Random – Direct Relationship										
Weak	*90*	2.00	0.095	89	90	91	92	92	93	93
	80	1.15	0.095	78	79	81	82	83	85	86
	70	0.60	0.095	67	69	70	73	74	76	77
	60	0.17	0.095	57	59	62	64	66	67	70
	50	-0.25	0.095	46	49	51	53	55	58	60
	40	-0.65	0.095	36	39	41	44	45	48	50
	30	-1.10	0.095	27	29	31	33	35	37	39
Strong	*90*	1.65	0.225	87	89	91	93	94	95	96
	80	0.85	0.225	75	79	82	85	88	90	92
	70	0.29	0.225	63	67	72	76	80	84	87
	60	-0.15	0.225	52	58	63	68	73	77	81
	50	-0.58	0.225	41	47	52	58	63	68	73

Table 4.4. (cont'd) Specification for missing NIHSS models, including model parameters and estimated documentation rates in each category of NIHSS.

Scenario		Model Coefficients		Estimated NIHSS Documentation Rates NIHSS Score Category						
Effect	*% Doc.*	*Intercept*	*Beta*	*0-2*	*3-5*	*6-10*	*11-15*	*16-20*	*21-25*	*25-42*
	40	-1.45	0.225	23	27	32	36	42	47	53
	30	-2.00	0.225	15	17	21	25	30	34	40
Missing Not at Random – Inverse Relationship										
Weak	*90*	2.46	-0.095	92	91	90	89	88	87	86
	80	1.64	-0.095	83	81	79	78	77	74	73
	70	1.10	-0.095	73	71	70	67	65	64	61
	60	0.66	-0.095	64	61	60	57	54	53	50
	50	0.24	-0.095	54	51	49	47	44	42	38
	40	-0.16	-0.095	44	42	39	37	34	32	30
	30	0.60	-0.095	33	31	29	27	25	24	22
Strong	*90*	2.85	-0.225	93	92	89	87	85	82	77
	80	2.02	-0.225	86	83	80	76	71	65	62
	70	1.47	-0.225	77	74	69	64	59	53	48
	60	1.01	-0.225	69	64	58	53	47	43	35
	50	0.58	-0.225	59	53	48	42	36	32	26
	40	0.17	-0.225	49	43	37	33	28	23	19
	30	-0.28	-0.225	38	33	28	24	20	16	14

Note: Shading represents the % rate of documentation at the patient-level determined by the specified missing data models (i.e. intercepts and beta coefficients).

Collectively, we simulated five models of NIHSS documentation, which includes one MCAR model and four MNAR models (direct-weak, direct-strong, inverse-weak, inverse-strong). Each model was repeated to illustrate seven overall rates of NIHSS documentation, 30% to 90% by 10%, and a fully documented dataset. Finally, to determine the impact of hospital volume, we modified hospital patient volumes as n=100, 300, and 500. In total, there were 5 missingness models x 8 documentation rates x 3 hospital volumes = 120 simulations with S=500 samples per simulation of N=100 hospitals. In each permutation of missingness pattern, documentation rate, and hospital volume, hospitals were identified as "observed" outliers from their estimated hospital random intercept, and were rank-ordered based on calculated risk standardized mortality rates (RSMR). The details on hospital outlier identification and RSMR profiling are outlined below.

Section 4 – Hospital Profiling Methodology

At this point, we have generated data sets with fully observed NIHSS and missing NIHSS data based on different mechanisms of missing NIHSS data. These datasets were then used to profile hospital-level RSMRs as is done in real-life datasets, by only including patients with complete documentation of NIHSS. We utilized the hospital profiling methodology employed by CMS to calculate hospital 30-day ischemic stroke RSMRs, which employs a hierarchical logistic regression model.[19,33] The hospital RSMRs were obtained as the ratio of predicted (*P*) to expected (*E*) mortality – or the P/E ratio – multiplied by the overall unadjusted mortality rate (~15% for 30-day ischemic stroke mortality). The numerator of the P/E ratio is the predicted mortality in each hospital, given its case mix and hospital-specific deviation in mortality (i.e. hospital random intercept). The denominator of the P/E ratio is the expected mortality in that hospital given the same case-mix if it had the mortality of the average hospital (i.e. hospital random intercept equal to 0).[19,32] Hence, the predicted number is the number of expected mortalities in that "specific" hospital.[52] A *P/E* ratio of >1 represents poorer hospital performance than expected, and a *P/E* ratio of <1 represents better hospital performance than expected. The *P/E* ratio was then multiplied by the overall 30-day mortality rate (15%) to produce the hospital RSMR, which was subsequently rank-ordered from lowest (#1) to highest (#100) in each simulation scenario.

Section 5 – Assessments of Profiling Accuracy Using the Simulated Data

The primary assessment of this study is to determine the accuracy of profiling (i.e. hospital RSMR rank order) under difference scenarios of missing data and hospital volume. Hospital RSMR rank-order is the primary method of profiling used in the CMS Hospital Value-

Based Purchasing Program (HVBP).[6] We determined accuracy in three different ways. First, we estimated the correlation between the true hospital rank-order (as defined during data generation) and the observed rank-order (as defined by estimated RSMRs). Spearman rank correlation and Pearson correlation coefficients were both estimated between the true and observed performance rank order in each scenario of missing NIHSS data. This approach estimates profiling accuracy on a continuous scale, as opposed to a binary categorization, which is done in the next assessment. Again, because we know the hospital's "true" performance, these correlations assess the validity of the RSMRs to accurately rank-order hospitals. These data were generated for each scenario of missing NIHSS data, stratified by hospital stroke volume.

Second, we assessed the accuracy based on the ability of the HLM to accurately identify high and low hospital performers on mortality. We defined high/low performing hospitals as being in the top or bottom 5th percentile of rank order (i.e. 10% high/low performer prevalence) and 20th percentile of rank order (i.e. 40% high/low performer prevalence). These categorizations of performance have been frequently used in previous research.[53,108,109] A hospital is considered a true high/low performing hospital if the rank-ordered, assigned random intercept is in the top/bottom 5th or 20th percentiles. We compared the true high/low performer status with the rank ordered RSMRs, which were similarly categorized.

Because we simulated "true" performance, we are able to calculate the sensitivity (Se), specificity (Sp), and predictive value positive (PVP) and negative (PVN) of the HLM to correctly identify high/low performers. Sensitivity represents the ability of the model to correctly classify a hospital as a high/low performer, given that it is in fact a true high/low performer. Specificity

refers to the models ability to correctly classify non-high/low performer hospitals, given that they are not high/low performers. The predictive value positive of the model represents the proportion of hospitals classified as high/low performers by the model which are known to be high/low performers. Conversely, the predictive value negative is the proportion of hospitals classified by the model as non-high/low performers which are known to not be high/low performers. These calculations (Table 4.5) were generated for each scenario of missingness and stratified by hospital stroke volume. We plotted the average Se, Sp, PVP and PVN over all 500 replications for each scenario of missing data, stratified by hospital stroke volume.

Table 4.5. Calculations for sensitivity (Se), specificity (Sp) and predictive value positive (PVP) and negative (PVN) for true vs. observed high/low performer classification.

Observed High/Low Performer Status†	***True High/Low Performer Status****		***Calculation***
	Yes	*No*	
Yes	True Positive (A)	False Positive (B)	PVP = [A / (A+B)]
No	False Negative (C)	True Negative (D)	PVN=[D / (C+D)]
Calculation	Se = [A / (A+C)]	Sp = [D / (B+D)]	

Note: High/low performers were defined as being in the top/bottom 5th percentile of rank-ordered performance or 20th percentile of rank-ordered performance

* Determined from assigned hospital random intercept in data generation step (i.e. true performance)

† Determined by the estimated hospital RSMR from the HLM (i.e. observed performance)

Lastly, we estimated the average absolute change in rank-order position relative to the hospitals true rank position in each scenario of prevalence and mechanism of missing NIHSS data. In each sample (S=500), we calculated the absolute difference between the true hospital ranking, and observed hospital ranking from the rank-ordered RSMRs in each scenario of missing NIHSS. Next, hospitals were categorized by quintile of their true hospital ranking (i.e. 1-20, 21-40, 41-60, 61-80, and 81-100). In each quintile, we calculated the average absolute difference between the true hospital ranking and observed hospital ranking for each scenario of

missing NIHSS data, averaged over the S=500 samples. We then plotted the absolute average difference between true and observed rankings for each prevalence and mechanism of missing NIHSS data, stratified by the true quintile ranking and hospital stroke volume (n=100, 300, and 500).

Aim 3 – Results

Accuracy of Hospital RSMR Rank-Order

The spearman rank correlations between true and estimated hospital performance can be seen in Figure 4.3. Note that the mechanism of missing NIHSS data did not have an important effect relative to the effect of sample size as dictated by hospital stroke volume and NIHSS documentation. In the low stroke volume hospitals, the Spearman rank correlation coefficient between assigned and estimated random intercepts was moderate at best (ρ=0.72) when NIHSS was fully documented. As documentation decreased, the correlation fell to between ρ=0.52 and 0.47, depending on the mechanism of missing NIHSS data. While some variation between mechanisms of missing NIHSS data was observed, at any given level of documentation the differences in correlations were at most 5% between the different mechanisms. In moderate stroke volume hospitals, correlation was as high as ρ=0.87, but also fell as documentation reduced. However, even at the lowest levels of documentation, there was moderate correlation between hospital random intercepts (ρ=0.70). Correlation between rankings was high (ρ>0.80) in most scenarios of missing NIHSS data in large stroke volume hospitals. Pearson correlation coefficients can be seen in Figure B.1, and were almost identical.

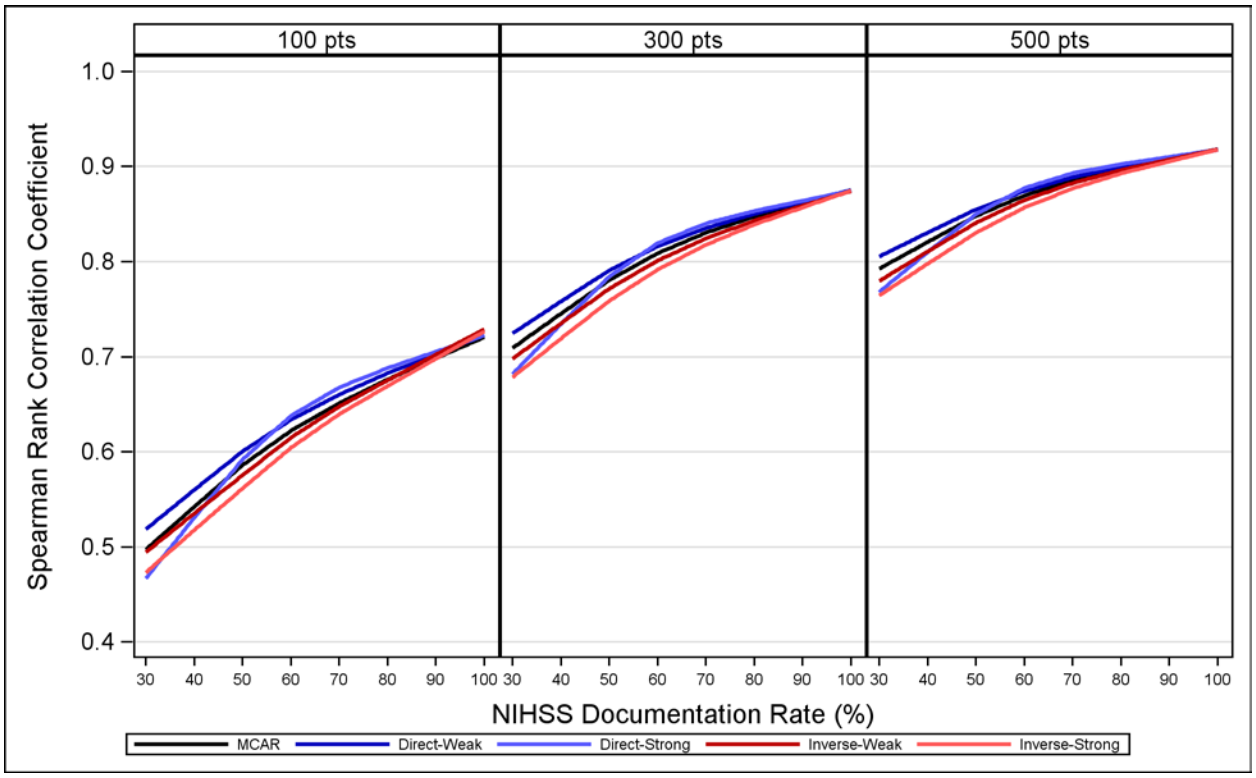

Figure 4.3. Spearman rank correlation coefficients between true rankings and RSMR rankings as NIHSS documentation increases under different mechanisms of missing NIHSS data. Results are stratified by hospital stroke volume.

Accuracy of High/Low Performer Classification

In general, as documentation increases, the number of true positives and true negatives increase, and the number of false negatives and false positives decrease (thus both Se and Sp increase). As hospital stroke volume increases, the number of true positives and negatives also increases, and the number of false positives and negatives decreases. There are no substantial differences in classification between mechanisms of missing NIHSS data.

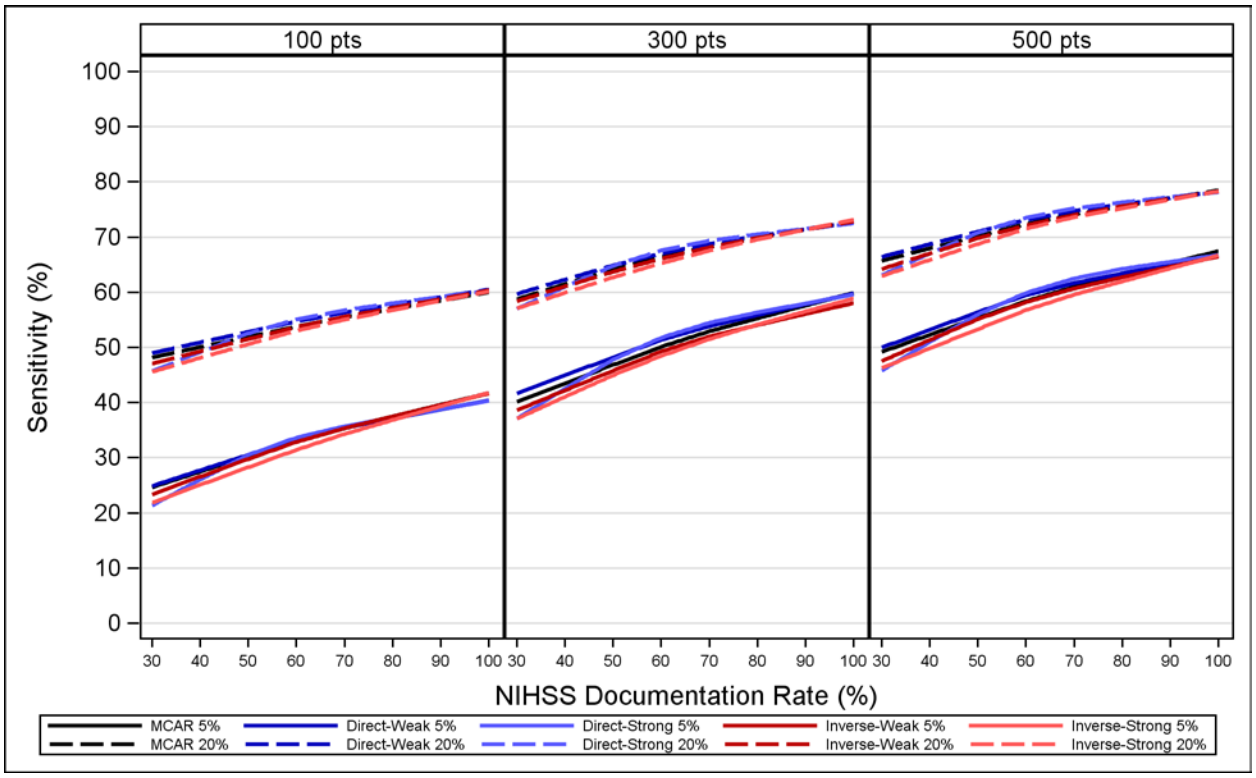

Figure 4.4. Sensitivity of HLM to classify hospitals as high/low performers based on top/bottom 5th (solid lines) and 20th (dashed lines) percentiles of mortality rank-order as documentation of NIHSS increases under different mechanisms of missing NIHSS data. Results are stratified by hospital stroke volume.

The sensitivity of the hierarchical logistic regression model to classify high/low performer hospitals according to estimated RSMRs, given that they are truly a high/low performer hospital can be seen in Figure 4.4. As documentation of NIHSS increases, sensitivity increases. Also, sensitivity was substantially higher when classifying high/low performing hospitals based on the top/bottom 20th percentiles compared to the top/bottom 5th percentiles. It should be noted that sensitivity is never greater than 80% in any scenario of missing NIHSS data or hospital volume. Notably, when documentation was complete in low volume hospitals, sensitivity was still worse compared to moderate and high volume hospitals

at the lowest levels of NIHSS documentation (30%). Differences in sensitivity between mechanisms of missing data was modest at each level of NIHSS documentation and hospital volume (<5%).

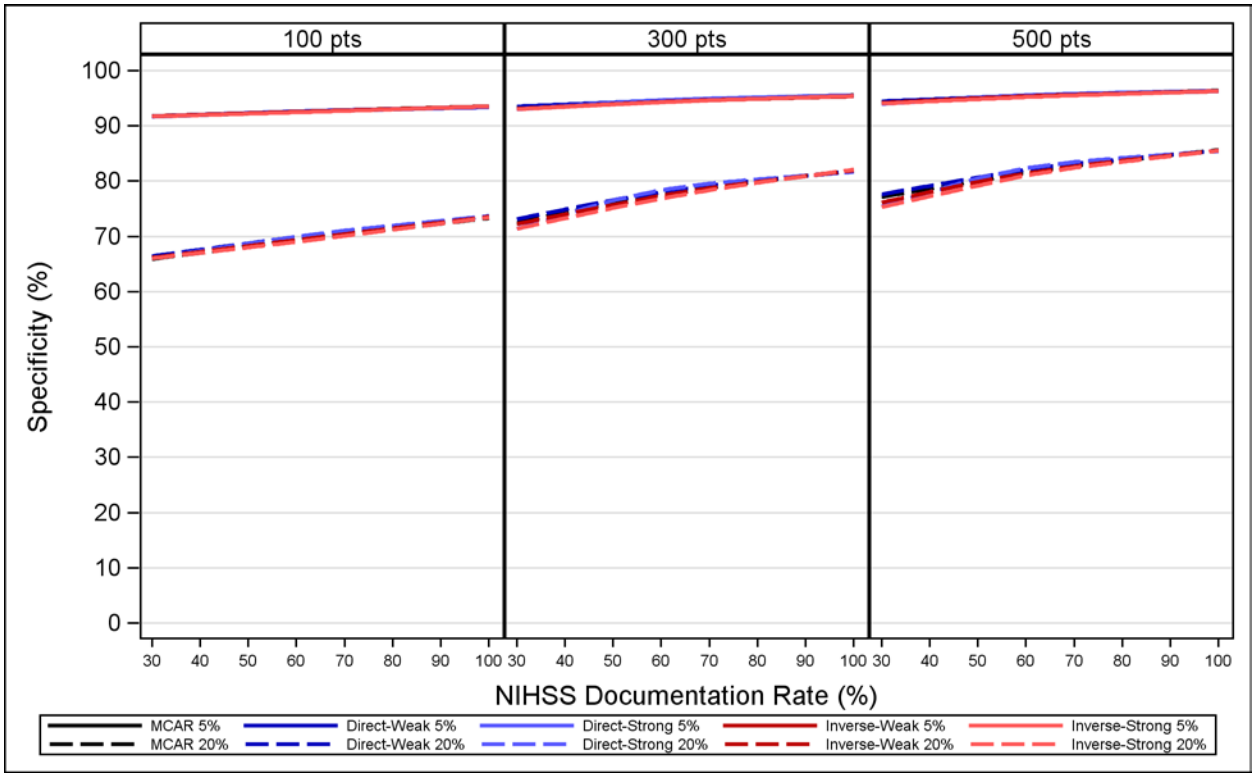

Figure 4.5. Specificity of HLM to classify hospitals as non-high/low performers based on top/bottom 5th (solid lines) and 20th (dashed lines) percentiles of mortality rank-order as documentation of NIHSS increases under different mechanisms of missing NIHSS data. Results are stratified by hospital stroke volume.

Figure 4.5 illustrates the specificity of the hierarchical model to identify non-outlier performing hospitals (i.e. not high/low performers). In contrast to sensitivity, the specificity of the HLM is much higher when classifying hospitals in the middle 90% (i.e., outliers are defined as the top/bottom 5th percentiles), and lower when using the middle 60% (i.e., outliers are defined as the top/bottom 20th percentiles. When classifying hospitals in the top/bottom 5th

percentiles, specificity was greater than 90% in all combinations of documentation and hospital volume, with only modest reductions as documentation fell. More substantial improvements in specificity were observed as documentation increased when hospitals were classified using the top/bottom 20th percentiles. Again, the mechanism of missing NIHSS data had little importance on specificity compared to the effect of sample size, as defined by hospital volume and NIHSS documentation. Although, differences between mechanisms were greater when classifying hospitals based on top/bottom 20th percentiles compared to 5th percentiles.

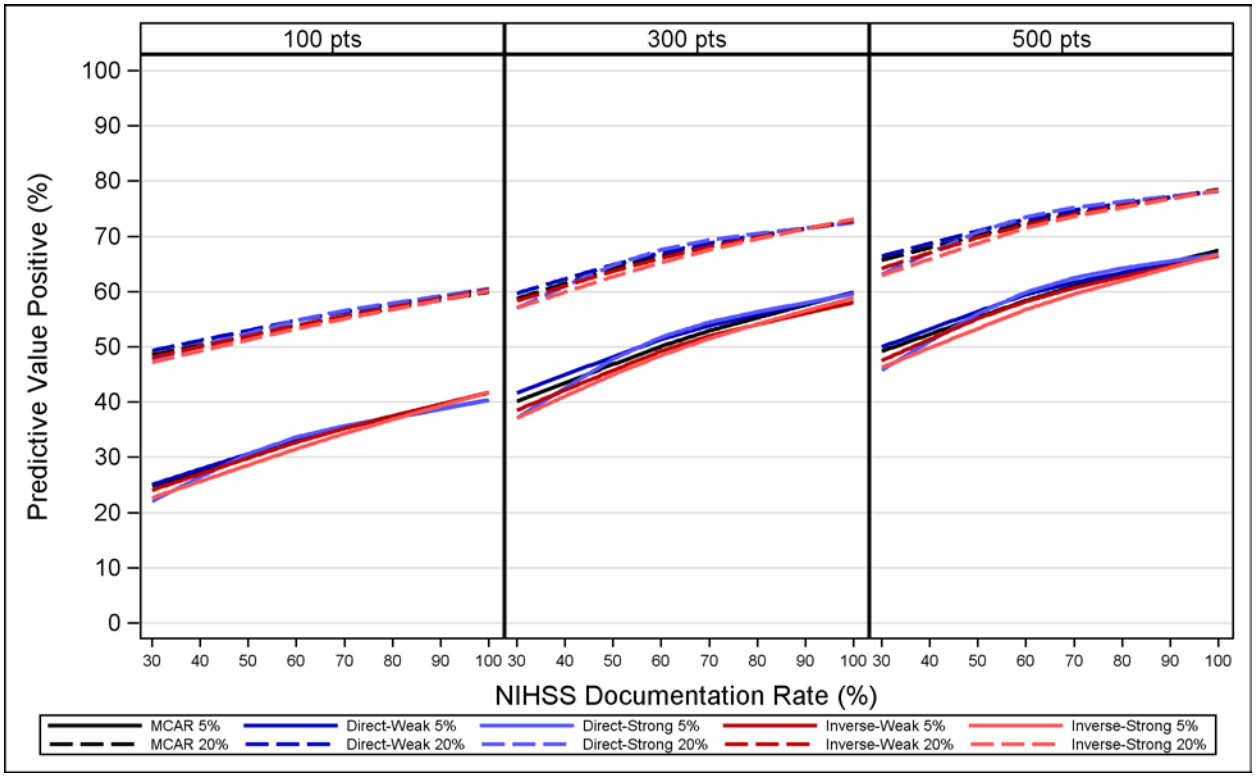

Figure 4.6. Predictive value positive of HLM to classify hospitals as high/low performers based on top/bottom 5th (solid lines) and 20th (dashed lines) percentiles of mortality rank-order as documentation of NIHSS increases under different mechanisms of missing NIHSS data. Results are stratified by hospital stroke volume.

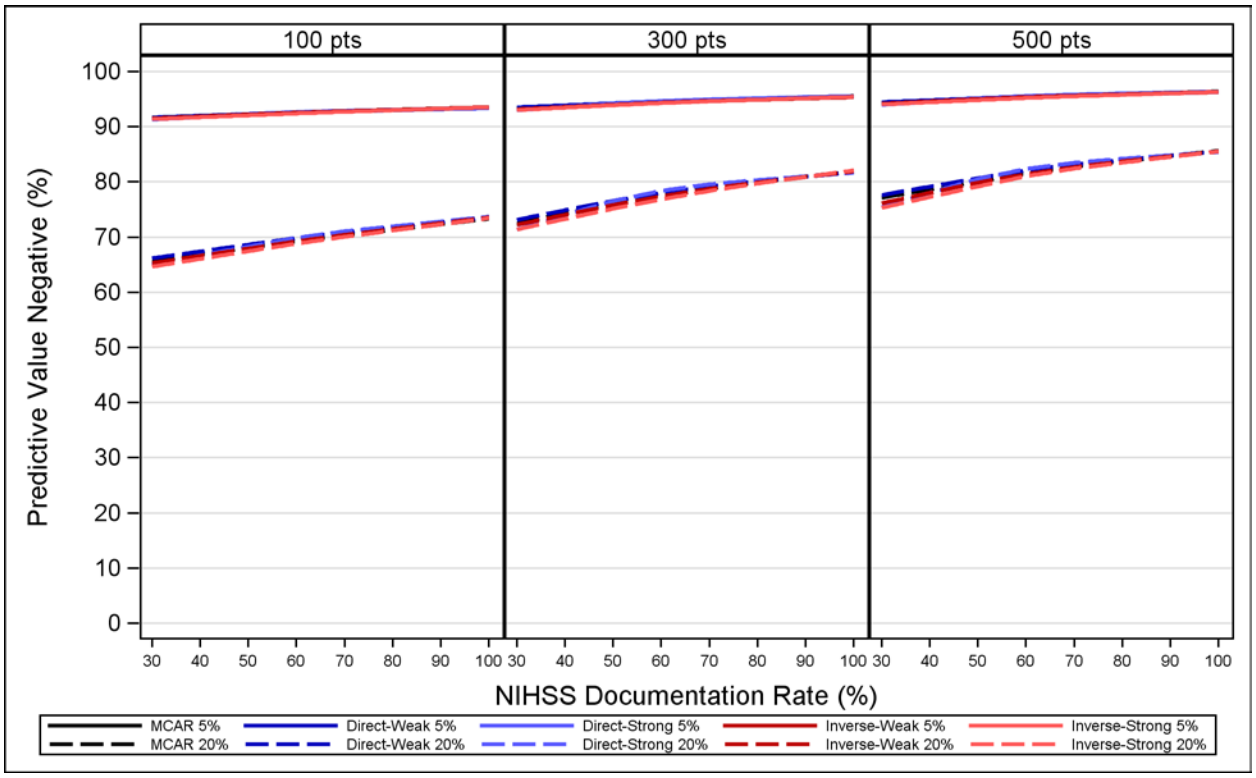

Figure 4.7. Predictive value negative of HLM to classify hospitals as non-high/low performers based on top/bottom 5th (solid lines) and 20th (dashed lines) percentiles of mortality rank-order as documentation of NIHSS increases under different mechanisms of missing NIHSS data. Results are stratified by hospital stroke volume.

Figures 4.6 and 4.7 show the predictive value positive (PVP) and negative (PVN) of the HLM to classify high/low performers, respectively. Patterns and values of PVP were similar to values obtained for sensitivity, because we categorized high/low performance based on rank-order cutoffs, the number of false positives and false negatives are essentially the same. The same can be said for the similarity between PVN and specificity. Briefly, PVP was greater when classifying hospitals based on top/bottom 20th percentiles compared to 5th percentiles, due to the greater prevalence of high/low performers. Consequently, PVN was lower when classifying hospitals based on top/bottom 20th percentiles compared to 5th percentiles. As documentation

of NIHSS increased, significant improvements in PVP and PVN were observed. PVP and PVN were highest in high volume hospitals, and lowest in low volume hospitals. Again, the mechanism of missing NIHSS data had modest impact on PVP and PVN.

We note that the average hospital high/low performer classification (i.e. true/false positives, true/false negatives) for each prevalence and mechanism of missing NIHSS, stratified by hospital stroke volume can be seen in Table A.1 (top/bottom 5th percentiles) and Table A.2 (top/bottom 20th percentiles).

Absolute Change in Hospital RSMR Rankings

Figure 4.8 shows the estimated magnitude of absolute change in observed rankings relative to the true (known ranking) stratified by the quintile of true hospital ranking. In general, the mechanism of missing NIHSS data did not have an effect, except at the lowest rates of NIHSS documentation. In low stroke volume hospitals, observed hospital rankings of the lowest and highest quintile of true hospital rankings changed as much as 25 positions on average when documentation was 30%. When documentation of NIHSS was complete, rankings of hospitals in the lowest and highest quintile still changed as many as 14 positions on average. It should be noticed that the results in Figure 4.7 are symmetrical in that they are the same for the 1st and 5th quintile, and 2nd and 4th quintile. Low volume hospitals in the second and fourth quintile of true ranking changed on average between 24 and 18 positions when documentation was 30% and 100%, respectively. Similar patterns were observed in moderate and large stroke volume hospitals, but the average change was smaller compared to low stroke volume hospitals. At most, moderate volume hospitals changed as many as 14 to 25 positions on average in the lowest and highest quintiles of true ranking when documentation was at 30%,

and only changed by 8 positions on average when NIHSS was fully documented. In large stroke volume hospitals, the average difference between true and observed hospital rankings was no more than 12 positions in the lowest and highest quintile in any scenario of missing NIHSS data.

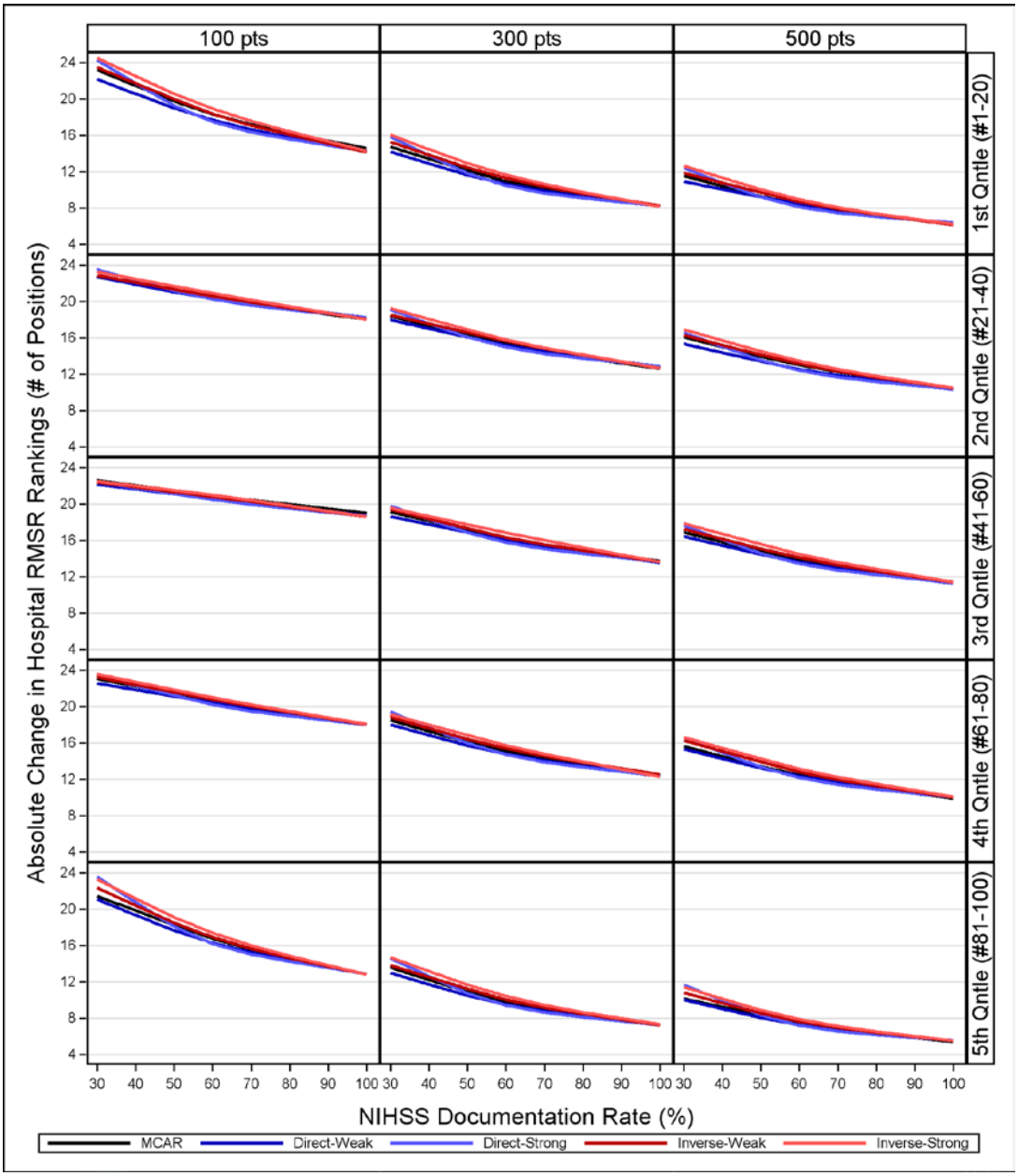

Figure 4.8. Average absolute change in hospital RSMR rankings (# of positions) as NIHSS documentation increases under different mechanisms of missing NIHSS data. Results are stratified by hospital size and quintile of true ranking.

Aim 3 – Discussion

In this study, we explored how current methods used to profile hospitals on ischemic stroke mortality are susceptible to inaccuracies when an important risk adjustment variable is missing. We imitated hospital-level rates of NIHSS documentation which are observed in the Michigan Stroke Registry, and theoretical mechanisms of missing NIHSS data motivated by previous analyses. To understand the importance of hospital stroke volume in our assessment, we conducted simulations with hospital stroke volumes of n=100, 300 and 500 ischemic strokes per hospital. Our main assessments were the ability of current methods to accurately rank-order hospitals according to their estimated risk standardized mortality rate (RSMR), to correctly classify high/low performing hospitals, and to estimate the average change in hospital RSMRs in the presence of missing data. Our primary finding was that the mechanism by which NIHSS was missing did not have a meaningful impact on the accuracy of hospital profiling per se, and was trumped by the much larger impact of the sample size that was determined by the level of NIHSS documentation and hospital size.

We hypothesized that when NIHSS documentation was associated with stroke severity, i.e. missing not at random (MNAR), the accuracy of hospital profiling would diminish compared to a missing completely at random mechanism (MCAR). On the whole, we found that the mechanism of missing NIHSS data did not lead to substantial differences in accuracy. Any observed differences in Se/Sp/PVP/PVN or in the correlation coefficients were less than 5% or $<\rho$=0.05, respectively, and the RSMR rank-order between mechanisms was less than 4 positions on average in any scenario of missing data. However, the mechanisms which were associated with inverse relationships (i.e. as NIHSS score increased, documentation decreased)

consistently had lower accuracy. This may be because under this assumption of missing data, the more severe patients are missing more frequently, and so the exclusion of these patients would lower the observed mortality in the hospital. As the rate of mortality and differences in mortality between hospitals reduce, accurate discrimination between hospitals becomes more problematic. The fact that we did not find the mechanism of missing NIHSS data to be very important could be explained by the modest variation in NIHSS between hospitals (ICC=2%), which were observed in the MSR. A necessary condition for a variable to have an meaningful effect in a risk adjustment model is that it should vary significantly between hospitals[26], and this has yet to be substantiated in regards to NIHSS. In Chapters 2, we found only modest differences in overall NIHSS at the hospital-level. If greater between-hospital variation in NIHSS were observed, the mechanism by which NIHSS is missing may play a larger role.

We found that reduced sample size – whether due to lower NIHSS documentation rate or low hospital case volume – resulted in poorer profiling accuracy, as depicted by substantially reduced rank correlation, and lower sensitivity and specificity. I hypothesize that changes in profiling accuracy based on sample size can be attributed, in part, to changes in the shrinkage of estimated random intercepts in the HLM model, which is inversely related to sample size.[110,111] Shrinkage is the phenomenon whereby estimated random intercepts in low volume hospital are "shrunken" toward the mean of all hospitals.[12,19,112] This is done because small volume estimates are presumed to be imprecise, and shrinkage accounts for the imprecision by stabilizing these estimates to the overall mean.[112] Because estimated random intercepts are utilized in calculating RSRMs, if there is greater shrinkage in low volume hospitals, subsequent RSMRs will also be "shrunken" toward the overall mortality rate.[19,109,111,113]

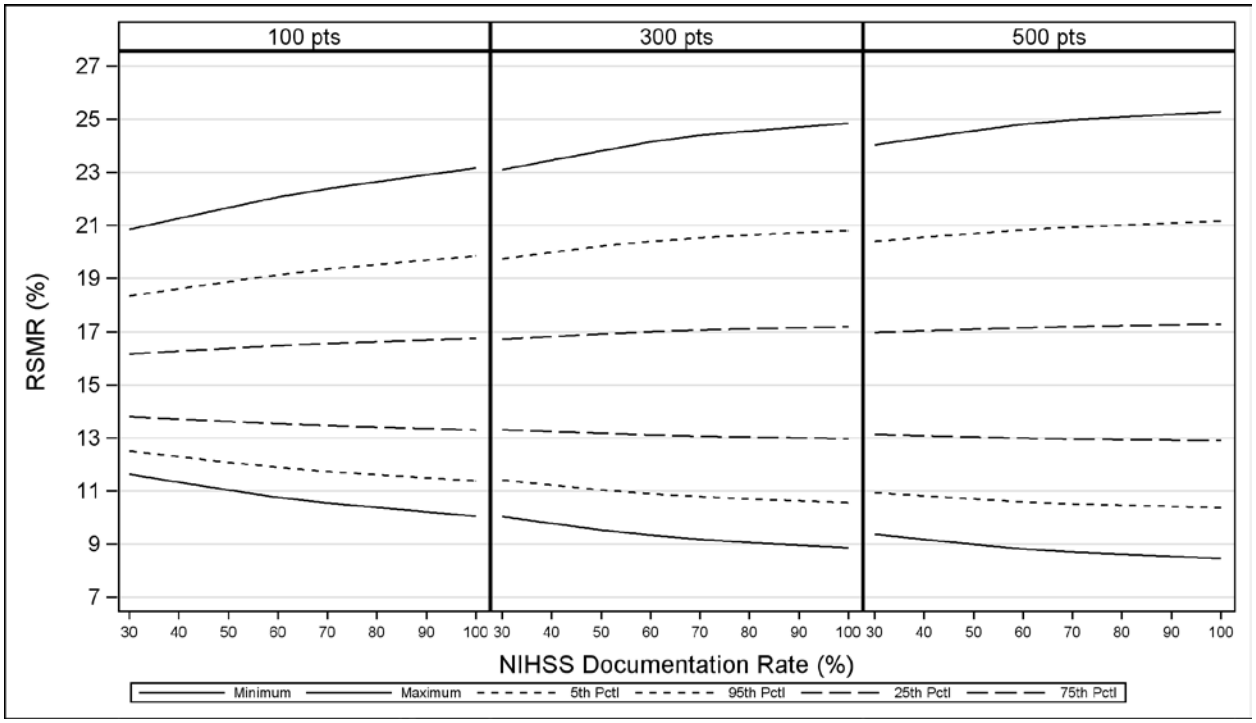

Figure 4.9. Illustrating the effect of shrinkage on RSMR distribution as depicted by range (i.e. minimum/maximum, solid lines), 5th/95th percentiles (dotted lines), and 25th/75th percentiles (dashed lines) of RSMRs. Estimates are the averages of 500 simulations for each of 100 hospitals.

To illustrate this phenomenon, we estimated the range (i.e. minimum, maximum), 5th/95th percentiles, and 25th/75th percentiles of estimated RSMRs for the 100 hospitals averaged over all S=500 samples. (Figure 4.9) The estimates are repeated for each scenario of NIHSS documentation (i.e. 30% to 100% by 10%) and hospital volume (n=100, n=300, and n=500 patients). (Figure 4.9) Because our previous findings did not support a significant role of missing NIHSS mechanism, here we only illustrate the MCAR mechanism. As sample size decreases, the plausible range of RSMR values decreases. Notably, while there are modest increases in the 25th/75th percentiles as documentation increases, there are much greater gains

in the observed range of RSMRs (i.e. minimum and maximum RSMRs) and 5^{th}/95^{th} percentiles. This illustrates the expansion of the RSMR distribution tails, indicating less shrinkage in estimated RSMRs.

We believe that shrinkage due to small sample size, either through NIHSS documentation rate or low hospital volume, is largely driving the reduced accuracy in RSMR profiling. Let's imagine that we are rank-ordering 100 hospital RSMRs, similar to our simulation methods. When sample size is small, the RSMRs for these hospitals will be more compressed around the overall mortality rate due to shrinkage. Any stochastic or random variability in these hospital RSMRs would lead to greater changes in profiling rank order, because they are more closely grouped together. Conversely, when sample size is large the same number of hospital RSMRs (n= 100) are less "shrunken", and so would be spread further apart. In this situation the same stochastic or random variability will be less impactful on RSMR rank order because they are more distanced apart. Thus, as sample size reduces, the accuracy of hospital performance profiling also reduces. A study by Silber et al. has illustrated the phenomenon of shrinkage in the context of Hospital Compare outlier performance by showing that the hierarchical model frequently underestimates poor performance in small hospitals with mortality rates moved close to the hospital average.[112]

Small sample size has long been a thorn in the side of hospital profiling.[112,114-116] Even when perfect risk adjustment is achieved, in typical clinical case volumes, much of the variation in performance measures is due to random noise, especially in centers with low volumes of cases (e.g. <100 annual ischemic strokes).[109] An oft cited benefit of the hierarchical model is its ability to produce more valid provider-specific estimates in low volume providers.[22,27] We

illustrate that even in the highest volume hospitals with complete documentation of NIHSS, the HLM approach still misses 2 of 10 hospitals in the top/bottom 5th percentiles of performance (Se=78%), and 3 of 10 (Se=68%) in low volume hospitals. If a more conservative definition of high/low performer is used (top/bottom 20th percentile), specificity in low volume hospitals becomes equally troubling, with more than 1 in 4 hospitals falsely identified as a high/low performing hospital (Sp=73%). The hierarchical model assumes that the variation in mortality left after adjusting for case-mix can be attributed to differences in hospital quality.[18] This study shows a substantial amount of random noise unrelated to true hospital performance influences hospital profiling. It is important to note that variation in our simulations cannot be attributed to confounding because our simulations achieved perfect case-mix adjustment. Until this noise can be accounted for, the accuracy of hospital profiling will remain suspect.

How we should interpret these findings relative to the current policies regarding hospital profiling methods is less clear. Low volume hospitals have frequently been shown to have poorer patient outcomes in ischemic stroke[117-119] and other clinical contexts.[120,121] As such, profiling methods should be robust enough to accurately capture performance outliers in small sample size scenarios. We also showed that as you expand the definition of high/low performers to include more hospitals, sensitivity and PVP is increased, but at the expense of reduced specificity and PVN. How you classify hospitals as high/low performers directly effects model sensitivity and specificity. The cost of identifying more false positives or false negatives depends on your viewpoint as a healthcare provider or consumer, and no correct answer exists.[122] If you are a patient or payer, such as CMS, it may be more beneficial to identify all the truly poor performing hospitals, at the risk of falsely identifying average or good performing

hospitals. On the other hand, hospitals may lose much needed financial reimbursements or be unfairly stigmatized if they are incorrectly labeled as a poor performer. Ultimately, both providers and consumers must be made aware of the limitations of current profiling methods to facilitate better interpretation of hospital profiling results.

There are some limitations and caveats to our study that should be considered. First, we simulated a 30-day mortality rate in our analysis, even though we did not have data on 30-day outcomes. However, current datasets which capture 30-day outcomes do not collect measures of stroke severity, so utilizing data with 30-day outcomes was not possible unless directly linked to administrative data. With ICD-10 codes set to include NIHSS, evaluation of hospital profiling methods using administrative data which includes both 30-day outcomes and stroke severity could be conducted in the future. Second, we did not obtain bootstrapped standard errors and 95% confidence intervals of individual hospital RSMRs to assess the accuracy of identifying statistical outliers, which is the approach used in Hospital Compare.[12] Future work should be done to test the accuracy of performance outliers using this method. Third, we did not compare our findings with the diagnostic ability of the current proposed CMS risk adjustment model, which is based on administrative data and does not include NIHSS.[18] A direct comparison would help illustrate the benefits and limitations regarding the current CMS risk adjustment model, and that of a model that includes NIHSS with various amounts of missing data. Fourth, while the models we specified to replicate missing NIHSS data were motivated by our analyses in Chapters 2 and 3, assessing the impact of missing data mechanisms rely on correct specification of the missing data model, which cannot be known with certainty. Fifth, in imputing the total risk score for individual patients we assumed a linear

relationship between the patient NIHSS component and non-NIHSS variables (i.e. the sub-risk score). This relationship may not be accurately captured, and should be validated using other data sources. Finally, our simulation parameters were based off a hospital sample which did not have substantial variation in severity between hospitals (ICC = 2.2%). Future studies should be conducted to assess how profiling accuracy is impacted when greater variation in stroke severity between hospitals is present, even though it remains unclear how much variation in severity actually exists.

In conclusion, the accuracy hospital profiling of ischemic stroke mortality is in large part a reflection of the sample size used to calculate hospital-level estimates, and sample size is influenced by both documentation rates of key risk adjustment variables and hospital case volume. Our simulation work shows that the mechanism of NIHSS missingness which is associated with severity (MNAR) has only a minimal impact on hospital profiling accuracy. However, even when NIHSS was completely documented, significant limitations in the accuracy of current methods used to profile hospitals should be acknowledged, especially in low volume hospitals. This study is innovative because it quantifies how much less accurate profiling becomes as missing data proliferates, and how accuracy interacts with hospital case volume. It also has advantages in that by using simulation methods we were able to determine the true ranking of hospitals performance with certainty and had no residual confounding by case mix.

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS

The overall aim of this study was to quantify the accuracy of hospital profiling when an important risk adjustment variable is missing. Specifically, using simulation based methods we investigated how hospital profiling based on ischemic stroke mortality is impacted when a strong predictor of mortality[56], stroke severity (i.e. NIHSS), is frequently undocumented.[53,54,56] Furthermore, we investigated how the mechanism by which NIHSS is missing impacts profiling accuracy, and how our findings are modified by hospital ischemic stroke volume. To test the underlying hypothesis that ischemic stroke patients with NIHSS documented are not a random sample of all patients, we conducted a series of analysis to identify patient- and hospital-level characteristics that are associated with NIHSS documentation in an existing clinical stroke registry (Michigan Stroke Registry). Additionally, we utilized the Heckman Selection Model as a diagnostic tool to assess the presence and magnitude of selection bias in the clinical registry.

Summary of Findings

Our analysis of the Michigan Stroke Registry (MSR) revealed a number of important findings. In Chapter 2, we found that at the patient- and hospital-level, patients with less severe stroke were less likely to have NIHSS documented. Beyond that, we found that documentation of NIHSS was a reflection of overall hospital-level documentation. Roughly a quarter of the variation in documentation was attributed to the hospital in which the patient was treated (ICC=25%). To illustrate the scale of this hospital-level variability, ICC's associated with hospital-level mortality and readmissions measures are typically below 5%.[23,24,123,124] This indicates that NIHSS documentation has both patient-level and hospital-level attributes; but

was not found to be accounted for by hospital characteristics such as annual stroke volume or Joint Commission primary stroke center status due to a lack of power at the hospital-level.

Notably, patients whose stroke symptoms had resolved by arrival to the ER had one tenth the odds of NIHSS documentation compared to patients who were still symptomatic upon arrival. Assuming that the absence of stroke symptoms is recorded with accuracy for these patients (who make up 6.5% of the registry) it would be reasonable to assume that these patients had an NIHSS of 0, which could be imputed into current registries with some confidence. We also found that patients who were administered tPA had higher rates of NIHSS documentation compared to non-tPA patients, which has been previously suggested.[66] If patients were not administered tPA because they missed the window for treatment, they may have worse outcomes compared to patients who received tPA. Thus, excluding these patients because they are missing NIHSS may also bias hospital-level estimates of mortality.

When we applied the Heckman Selection Model to the same MSR data in Chapter 3, we found as expected, evidence of selection bias in patients with documented NIHSS, although it was rather modest (correlation coefficient: $\rho = 0.11$). The positive correlation also indicates that as NIHSS increases (i.e. strokes are more severe), the probability of NIHSS documentation also increases. We repeated the analysis in time periods with lower (documentation = 67% in 2009-2010) and higher rates (documentation = 87% in 2010-2012) of NIHSS documentation to assess the impact of missing NIHSS data prevalence. Selection bias increased marginally when documentation was lower; and conversely, when documentation was higher, selection bias decreased. Together, these analyses support the hypothesis that patients with documented

NIHSS are not simply a random sample of all stroke patients at the patient- or hospital-level, and that subsequent hospital-level estimates using this sample may subsequently be biased.

How selection bias at the patient-level translates to bias in hospital-level estimates is not clear, and was what originally motivated our study. We employed computer simulations to estimate the accuracy of hospital profiling based on ischemic stroke mortality under various mechanisms and prevalences of missing NIHSS data. Simulations were essential in this instance, because they allow us to assign a known (true) hospital-level mortality performance, which is impossible to determine in real-world conditions.[29,82,108,109] Since true hospital performance is known, we can measure the diagnostic accuracy of profiling, using measures of as sensitivity, specificity, and predictive value positive and negative, by comparing true hospital performance with the performance estimated using current hospital profiling methods under various scenarios of data documentation.

There are some other benefits to computer simulations that should be noted. One benefit is that we employed risk adjustment models which were not subject to inadequate risk-adjustment.[108,109] This is because the fitted risk adjustment model was identical to the model used in the data generation process. Consequently, any hospital misclassification cannot be attributed to residual confounding from unmeasured case-mix differences, but to random variation. Simulations also allow one to explore a variety of scenarios to be developed in order to explore the modifying effect of other variables (such as hospital volume) and are ideal to conduct sensitivity analyses of underlying parameters and assumptions.[125] But, simulation studies can be difficult to understand, which can lead to confusion when interpreting results

and making correct conclusions.[126] They also rely on correct assumptions about real-world data, which should be justified at each step.[125]

The results from our simulation studies in Chapter 4 can be succinctly summarized as follows: 1.) the mechanism by which NIHSS is missing (i.e., MCAR, MNAR) plays only a minor role in the accuracy of profiling, 2.) because of its effect on sample size the NIHSS documentation rate (where cases with missing NIHSS data are deleted) has a substantial impact on the accuracy of profiling, and 3.) the relationship between NIHSS documentation and profiling accuracy was exacerbated by hospital ischemic stroke volume. In sum, the *mechanism* by which NIHSS is missing is not as important in the context of profiling accuracy as is the *amount* that is missing, and in the *size of hospitals* in which it is missing. This study illustrates fundamental limitations of the profiling method by showing how the underlying sample size has a profound effect on the accuracy of performance profiling.

The first assessment in Chapter 4 was the ability of the hierarchical model to accurately estimate hospital rank order. We compared the rank order of true hospital performance to the estimated rank order generated from the RSMR estimates. This assesses the accuracy of profiling on a continuous scale, as opposed to the subsequent assessments, which dichotomized hospitals as either outliers (i.e., high/low performers) or not outliers based on arbitrary cut points. We found that in moderate and high volume hospitals, correlation between these true and observed ranking was generally high (>0.80). But, as documentation of NIHSS decreased, correlation between rankings also decreased, more markedly in moderate sized hospitals. With perfect documentation, correlation between rankings in low volume

hospitals was moderate (ρ=0.72), but dropped to almost ρ=0.50 when documentation reduced to 30%. The mechanism of missing NIHSS had negligible effect on the correlation coefficients.

The next assessment in Chapter 4 was the ability of the hierarchical logistic model to correctly classify high/low performing hospitals, based on the true and estimated performance rank-order. Two definitions of outlier hospitals were used; top/bottom 5th percentile and top/bottom 20th percentile hospitals. We found that, in general, as documentation of NIHSS reduced, the model sensitivity, specificity, PVP, and PVN all reduced. There was little variation in these measures between mechanism of missing NIHSS at a given level of documentation and hospital volume. Sensitivity was never higher than 80% in any scenario, and as expected was much higher when categorizing hospitals in the top/bottom 20th percentiles compared to top/bottom 5th percentiles, because it is easier to classify hospitals as high/low performers when it is defined more broadly. Conversely, specificity was much higher when categorizing hospitals into the top/bottom 5th percentiles. Again, the mechanism of missing NIHSS data had only modest effects. Similar effects were observed for PVP and PVN

Our final analysis in Chapter 4 was to assess the magnitude of change between true performance rankings and rankings based on calculated hospital risk-standardized mortality rates (RSMRs). We found that observed performance rank order (which ranged from 1 to 100 in each simulation) could change significantly compared to the true performance rank order, and this was especially evident in low volume hospitals. Changes in rankings between different mechanisms of missing NIHSS data were again only modest or almost non-existent. Even with perfect NIHSS documentation and perfect case-mix adjustment, hospitals in the top and bottom quintile of true performance rankings changed on average 13 positions. As documentation of

NIHSS reduced, the average difference between observed and true performance rank order increased to almost 24 positions in low volume hospitals in the top (1-20) and bottom (81-100) quintile of hospital true performance rankings. While changes in position were not as volatile in moderate and high volume hospitals, hospitals still changed at least an average of 5 positions in the top and bottom quintile of true performance rankings.

Again, these findings illustrate that random noise after risk adjustment negatively impacts hospital profiling, especially when sample size is low, due to shrinkage of RSMR point estimates towards the mean. Previous work in the GWTG-Stroke population showed that including NIHSS in risk adjustment improved the model fit and reclassified a significant proportion of hospitals.[53] However, more than half of ischemic stroke patients in GWTG-Stroke were excluded from this analysis because they did not have NIHSS documented. We showed that at this rate of NIHSS documentation, hospital RSMR rankings could change on average 9-16 positions in high volume hospitals, 12-18 positions in moderate volume hospitals, and 20-23 positions in low volume hospitals due to random variation alone. Given the great degree of inaccuracy at this level of reporting, significant changes in rankings are not unexpected.

Limitations

There are several limitations of this study. First, our analysis used the Michigan Stroke Registry (MSR), which has data on a limited number of hospitals and may not be representative of all stroke patients. A greater proportion of MSR patients go to teaching hospitals (93% vs. 61%) and Joint Commission primary stroke center hospitals (78% vs. 65%) compared to patients in the national GWTG-Stroke registry.[96] Thus, patients in the MSR may be more similar to each other compared to what may be seen in the GWTG-Stroke registry, and are likely different than

patients treated at all US hospitals. A repetition of our simulations using parameters estimated from a more comprehensive dataset, such as the national GWTG-Stroke registry data linked with Medicare claims data, would be useful in generalizing our results to data used in CMS pay-for-performance schemes. Access to Medicare claims data would also allow for a direct comparison with the risk-adjustment model currently proposed to profile hospitals on ischemic stroke 30-day risk standardized mortality, which was not done in this study.[18] Linking Medicare claims data to GWTG-Stroke registry data may also allow for an evaluation of both a model with and without NIHSS on the proposed 30-day risk standardized readmission measure for ischemic stroke.

With regard to our simulations, there are other limitations to consider. First, we simulated variation in patient- and hospital-level risk of mortality which reflects data observed in the MSR. However, this variation was not substantial (ICC = 2.2%), and may not reflect what is observed in most hospitals. Although this between-hospital variation in mortality is small, it is consistent with prior estimates in the literature which are typically <5%.[23,24,123,124] Additional simulations should be conducted to reflect a greater between-hospital variation in risk, which may have important consequences on our findings. We also did not examine the accuracy of profiling as reported by the Hospital Compare program, which identifies hospitals with better- or worse-than-expected mortality rates based on a statistical test of the estimated RSMRs relative to the average hospital.[12] Future work should examine how missing data impacts the accuracy of statistical outlier identification as used by the Hospital Compare program. However, a previous study has already shown that the methods used in Hospital Compare to identify outlier hospitals significantly underestimates poor performance in low volume hospitals

due to the shrinkage phenomenon.[112] Furthermore, the missing not at random (MNAR) mechanisms used in simulations were motivated by findings in Chapters 2 and 3, but may not represent the actual missing data mechanism. Additional mechanisms, such as bimodal mechanisms or mechanisms related to other important covariates, should be explored to complement our analyses. However, our findings suggest that the mechanism by which data is missing may have minimal impact on performance profiling.

Including NIHSS in Risk Adjustment Models for Stroke Performance Measures

Advocates for including NIHSS in risk adjustment models for ischemic stroke performance measures will be energized by its addition to ICD-10 coding in administrative data.[127] Given its importance in patient-level outcome prediction[56], the enthusiasm is warranted. However, including it in risk adjustment models for hospital-level estimates of performance should be approached with caution because it is frequently undocumented in clinical registries. How complete documentation of NIHSS will be in ICD-10 is unknown. But, documentation of NIHSS has been improving in clinical registries, such as the Get With the Guidelines – Stroke national registry, where in recent years it has been as high as 70%. It is likely that hospitals participating in clinical registries such as GWTG-Stroke represent a more engaged and trained subset of hospitals and a concerted effort has been made by the GWTG-Stroke program to improve NIHSS documentation in participating hospitals. Hence, it may be unreasonable to expect that NIHSS documentation in hospitals not involved in such programs would achieve similar levels as those seen in more recent years of the GWTG-Stroke. Since our study has shown that hospital-level documentation of NIHSS is a significant driver of patient-level documentation, and has tremendous impact on the accuracy of ischemic stroke hospital

profiling, eagerness to include NIHSS in risk adjustment should be tempered, until NIHSS has increased to an acceptable level, such as 80% or greater.

Our findings also showed that hospital-level NIHSS did not vary sufficiently between hospitals in our sample. Little hospital-level variation of NIHSS was also illustrated in a study of VA hospitals.[58] The rational for the addition of NIHSS as a risk adjustment variable is weakened if it does not vary sufficiently between hospitals to warrant inclusion.[26] However, both the VA study and our study sample may not be representative of most hospitals. While it has been suggested that hospitals which see more severe strokes – such as tertiary referral centers or Joint Commission primary stroke centers – may be at greater risk for misclassification if stroke severity is not included in risk adjustment[26,48,59], little evidence has been presented to support that claim.

Further research should be done to investigate the amount of between-hospital variation in stroke severity. Since sufficient between-hospital variation in patient-level variables is a prerequisite for inclusion in risk adjustment models, understanding the extent of between-hospital variation may help guide decisions about the need to include stroke severity in models for ischemic stroke mortality and readmissions. Analysis should also be done to understand if variation is driven by hospital-level characteristics, such as tertiary referral centers or Joint Commission primary stroke centers. These characteristics may be able to serve as proxies for stroke severity, which are easier to obtain than measures of stroke severity on every patient.

Simulation studies could be used to assess how modifying the variation in case-mix at the hospital level – particularly as it pertains to stroke severity – improves the accuracy of

hospital profiling. The variation in case-mix in our simulations reflected observed differences in the MSR, but altering parameters of our simulation would allow us to investigate the impact of greater variation in case-mix between hospitals. This could be achieved in two ways: 1.) by increasing the overall amount of variation in hospital-level case mix, and 2.) by increasing the proportion of case-mix which can be attributed to the hospital-level (i.e. intraclass correlation of case-mix). This analysis would illustrate how the presence of missing NIHSS data impacts hospital profiling when greater disparities in case-mix between hospitals are present.

Critique on Current Profiling Methodologies

The analysis presented here highlights important drawbacks to current methods of hospital profiling in general. Pay-for-performance models assume that profiling methods can accurately compare hospitals on predetermined performance measures after accounting for patient case-mix. [19,20] However, a growing body of literature suggests that current profiling methods are inadequate. Low sample size is a well documented limitation of hospital profiling[114-116], which is especially problematic in the context of stroke, given that low volume settings have been shown to have higher rates of mortality in ischemic stroke[117-119] and in other clinical applications, such as surgical outcomes.[120,121,128] Simulation studies have found that the accuracy of hospital report cards in case volumes typically seen in clinical settings is low, and further deteriorates in lower case volume hospitals.[108,109]

Our analysis echoed these concerns, showing that profiling accuracy is inextricably linked with provider sample size. By any measure of accuracy, the estimated RSMR from the HLM used to profile hospitals is less accurate as sample size reduces, either through hospital volume or missing data. This is due to the effect of shrinkage in RSMR estimates toward the

mean when sample size is small, which was illustrated in Figure 4.9. Even when documentation of NIHSS is 100%, our simulations show serious limitations in the accuracy of current profiling methods. Data in Table 5.1 illustrate the observed Se, Sp, PVP and PVN across the hospital volumes under the scenario of complete NIHSS documentation. Even in these best case scenarios, when the definition of high/low performers is strict (i.e. top/bottom 5th percentiles), sensitivity and PVP are quite poor, while specificity and PVN are generally high. If the definition of top/bottom performer is expanded to include more hospitals (i.e. changed from top/bottom 5th percentile to 20th percentile), sensitivity and PVP increase, but at the expense of specificity and PVN.

Table 5.1. Diagnostic ability of hierarchical logistic model to identify hospital high/low performers when documentation of NIHSS is complete (i.e. no missing NIHSS data), stratified by definition of high/low performer and hospital stroke volume.

	Top/Bottom 5th Percentiles			***Top/Bottom 20th Percentiles***		
	Hospital Stroke Volume					
Diagnostic Measure	*n=100*	*n=300*	*n=500*	*n=100*	*n=300*	*n=500*
Sensitivity	41%	58%	67%	60%	73%	78%
Specificity	94%	95%	90%	73%	82%	85%
Predictive Value Positive	42%	59%	67%	60%	73%	78%
Predictive Value Negative	94%	95%	96%	73%	82%	85%

Note: Diagnostic measures calculated using data from Table A.1 (top/bottom 5th percentiles) and Table A.2 (top/bottom 20th percentiles)

These findings illustrate the need to apply optimal decision making theory in order to guide hospital performance profiling benchmarks by placing relative values/costs on identifying false positive vs. false negative high/low performers.[129,130] By making the definition of high/low performer more conservative, i.e. changing from top/bottom 5% to 20%, we substantially improved the sensitivity and PVP of HLM, but at the cost of specificity and PVN.

Austin, et al. found that decisions about the significance-level used to classify hospitals as performance outliers can lead to outlier designations which are more or less preferable to patients as opposed to providers, based on the values associated with false positive or false negative hospital outliers.[122] Decisions should be made regarding which classification is more important, false negatives or false positives, and the potential economic impact of these decisions in the context of pay-for-performance incentive structures.

Given that low hospital case volume will be omnipresent in any hospital profiling scheme, future research should explore solutions to improve the accuracy of performance profiling in these hospitals. In addition to traditional frequentist methods, Bayesian methods can be used to provide further evidence that a hospital may indeed be a performance outlier.[20,122,131,132] Longitudinally profiling hospital performance may also be useful, especially as data collection in hospitals becomes routine.[133] Further simulation studies may also provide insight into how many years of data should be pooled to profile hospitals accurately, especially in the case of low volume hospitals.

Although we used data from clinical registries for the simulations, our findings are equally relevant when considering the use of administrative data to profile hospitals. Administrative data has previously been shown to lack important prognostic indicators as compared to clinical datasets.[22,134-136] Even if risk adjustment models are developed in administrative datasets with similar model fit compared to clinical models, coding and documentation inconsistencies between hospitals can threaten the validity of the hospital-level estimates.[40] Krumholz, et al. outlined standards for using administrative data to profile hospitals, which outlines that data must be sufficiently high-quality and timely.[22] We illustrated

how hospital-level performance profiling measures can be impacted when data are not sufficiently high quality, such as in the case of NIHSS. A greater emphasis should be placed on the quality and completeness of data used in risk adjustment models, whether data from administrative sources or clinical registries are used.

If current profiling methods are going to continue to use administrative data, solutions to missing or miscoded data are needed. Missing data methods such as maximum likelihood estimation or multiple imputation may provide a solution to address frequently undocumented or miscoded data. Multiple imputation has been shown to facilitate the identification of provider outlier status, but can be sensitive to the assumptions made about reasons for missing data.[75] Our results show that the amount of missing data is more problematic than the mechanism by which it is missing. Thus, any bias associated with imputing values using missing data methods, especially when data are MNAR, is probably outweighed by the gain in sample size. Future research could use simulation methods to compare the accuracy of hospital-level estimates generated using imputation of missing data to those calculated by a complete case analysis that excludes observations with missing data.

As mentioned in Chapter 1, mortality-based performance measures of mortality already suffer from a plethora of limitations, including the inability to accurately discriminate between "good" and "bad" hospitals[39], the sensitivity of RSMRs to risk adjustment model specification[41-43], and that few deaths in hospitals may actually be preventable.[44,45] Additionally, low variability in hospital-level performance measures between providers has been also shown to reduce the accuracy of hospital performance classification.[137] In Hospital Compare – the Medicare hospital performance reporting system[12] – hospitals in the top or bottom tier of performance have been

shown to be not statistically different from at least one hospital in the middle tier of performance, suggesting that side-by-side comparisons using publicly reported performance profiling measures may be misleading to consumers [138] Our research extends this observation by illustrating that profiling accuracy is quite low in many instances, and that calculated RSMRs are subject to a substantial amount of random noise, especially when sample size is low, either through low hospital volume or through missing data.

Future Directions

We proposed a number of future directions as a result of our study, which can be summarized as follows. First, a better understanding of hospital-level variation in stroke severity is needed to assess its utility as a risk-adjustment variable for hospital-level performance measures. Simulation studies may also illustrate how much between-hospital variation in stroke severity (leading to case mix differences) is needed to impact hospital profiling accuracy. Altering parameters in our simulations to mimic increased case-mix variation is also needed to understand its impact on our findings. Second, obtaining simulation parameters from a more comprehensive dataset, such as the GWTG-Stroke registry, would improve the generalizability of our findings. Linking this with claims data would allow us to compare an NIHSS risk-adjusted model with the current CMS model, obtain simulation parameters associated with 30-day outcomes, and evaluate the accuracy of the 30-day readmissions measures (RSRR) as well. Third, if NIHSS is to be included in profiling, missing data methods (e.g. multiple imputation) should be explored to address the problem of missing data, and its impact on profiling accuracy. Fourth, simulations should be done to obtain bootstrapped standard errors and 95% confidence intervals for estimated 30-day outcomes of

individual hospitals, to evaluate the accuracy of the CMS method to identify performance outliers as currently employed by the Hospital Compare program.[12] Finally, we should explore other statistical methods such as decision making theory to guide outlier performance categorization, and Bayesian and longitudinal methods to assess their utility to accurately profile hospitals compared to the current HLM method.

Conclusion

In sum, there are significant concerns about the validity and reliability of current profiling methods which should be considered when developing policies which rely on accurate performance comparisons. But, in spite of this evidence, healthcare stakeholders, such as the CMS, are doubling down on pay-for-performance models which are tied to performance profiling. U.S. Secretary of Health and Human Services Sylvia Burwell announced that by 2018, 90% of Medicare fee-for-services payments will be tied to quality or value[139], which emphasizes that it is critical to have reliable and valid measures of quality and value. Unless methods to compare hospital performance are improved, a substantial proportion of hospital could be unfairly punished for poor performance that may not actually be poor (i.e. low predictive value positive), and hospitals that are providing poor care may go undetected (i.e. low sensitivity or predictive value negative). It is important to note that we are not advocating for the abandonment of pay-for-performance models or hospital profiling, but simply suggesting that intrinsic limitations to current methods should be realized, and further research should be conducted, such as in the outlined in the future directions section above, to create more robust profiling methodology.

Ultimately, hospital performance profiling should be one method in a larger repertoire of tools to assess hospital quality of care. Healthcare is multidimensional and interdependent, and excelling in every category of hospital quality is important in its own right. While statistical methodologies used to assess healthcare quality should continued to be improved upon, healthcare providers should strive to improve all aspects of care, rather than focusing on a handful of quality measures. Thomas H. Lee astutely conveyed this notion in a recent editorial[140] when he stated: "Reliability matters. Safety matters. Efficiency matters. Patient experience matters. All of these dimensions of performance are intertwined, and interact to define the quality of an institution's care."

CHAPTER 6: SUMMARY

Pay-for-performance schemes, which are currently used as a model to improve the quality and value of care, rely on accurate comparisons of hospital performance, i.e. hospital profiling. Proposed measures to profile hospitals on ischemic stroke mortality and readmissions at thirty days have been controversial because they lack a measure of stroke severity. The National Institutes of Health Stroke Scale (NIHSS) is a commonly used measure of stroke severity that is highly predictive of patient outcomes; however, it is frequently missing in large scale clinical databases and is currently completely absent from administrative data. With the announcement that NIHSS is to be included in ICD-10 administrative coding, there will be pressure to include it in risk adjustment models for ischemic stroke outcomes. But, if the subsample of patients with documented NIHSS is a biased sample of ischemic stroke patients, there is the potential that hospital-level estimates of mortality may also be biased, but the extent of which is unknown.

The main contribution of this study is a quantification of the impact that missing data on an important risk adjustment variable has on the accuracy of hospital profiling. We conclude that the accuracy of hospital profiling is strongly impacted by missing data, although this is not because the mechanism by which missing data occurs is important. Rather, missing NIHSS data has an important effect on profiling because it results in a smaller "effective" hospital sample size, which has a much stronger effect on profiling accuracy due to the impact of shrinkage on estimated hospital random intercepts. Moreover, this study also illustrates limitations of current profiling methods, even when perfect documentation and risk adjustment are achieved. It is noteworthy that documentation of NIHSS is driven by a combination of both patient-level

and hospital-level factors. Furthermore, we note that hospital-level variation in actual NIHSS scores in our sample of hospitals is not substantial, which, if true, would lessen the rationale for including NIHSS in risk adjustment models. These findings are important when considering covariates to be used in risk adjustment models, as well as the validity of profiling hospitals on ischemic stroke mortality, and other hospital-level performance measures.

APPENDICES

Appendix A: Supplementary Tables

Table A.1. Average proportion (%) of hospital high/low performer classification for top/bottom 5th percentile of rank-order (true positive, false positive, true negative, false negative) for different mechanisms of missing NIHSS data, stratified by hospital stroke volume (*n*=100, 300, and 500).

Mechanism of Missing NIHSS	***NIHSS Doc. Rate***	***Average Hospital High/Low Performer Classification (%)***											
		n=100				*n=300*				*n=500*			
		TP	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*
MCAR	*30*	2.3	7.7	82.4	7.6	3.9	6.1	83.9	6.1	4.8	5.2	84.8	5.2
	40	2.8	7.2	82.8	7.2	4.4	5.6	84.4	5.6	5.2	4.8	85.2	4.8
	50	3.2	6.8	83.2	6.8	4.7	5.3	84.7	5.3	5.6	4.4	85.6	4.4
	60	3.3	6.7	83.3	6.7	5.1	4.9	85.1	4.9	5.9	4.2	85.9	4.2
	70	3.5	6.5	83.5	6.5	5.3	4.7	85.3	4.7	6.1	3.9	86.1	3.9
	80	3.8	6.2	83.8	6.2	5.6	4.4	85.6	4.4	6.4	3.6	86.4	3.6
	90	3.9	6.1	83.9	6.1	5.8	4.2	85.8	4.2	6.5	3.5	86.5	3.5
	100	4.2	5.8	84.2	5.8	5.9	4.1	85.9	4.1	6.7	3.3	86.7	3.3
MNAR Direct – Weak	*30*	2.4	7.6	82.5	7.5	4.1	5.9	84.1	5.9	4.9	5.1	84.9	5.1
	40	2.8	7.2	82.8	7.2	4.5	5.5	84.5	5.5	5.4	4.7	85.4	4.7
	50	3.1	6.9	83.1	6.9	4.9	5.1	84.9	5.1	5.7	4.3	85.7	4.3
	60	3.3	6.7	83.3	6.7	5.2	4.8	85.2	4.8	5.9	4.1	85.9	4.1
	70	3.6	6.4	83.6	6.4	5.4	4.6	85.4	4.6	6.3	3.7	86.3	3.7
	80	3.7	6.3	83.7	6.3	5.6	4.4	85.6	4.4	6.3	3.7	86.3	3.7
	90	3.9	6.1	83.9	6.1	5.8	4.2	85.8	4.2	6.5	3.5	86.5	3.5
	100	4.0	6.0	84.0	6.0	5.9	4.1	85.9	4.1	6.7	3.4	86.7	3.4
MNAR Direct – Strong	*30*	2.0	8.0	82.5	7.5	3.7	6.3	83.7	6.3	4.5	5.5	84.5	5.5
	40	2.6	7.4	82.7	7.3	4.2	5.8	84.2	5.8	5.1	4.9	85.1	4.9
	50	3.3	6.7	83.3	6.7	5.0	5.0	85.0	5.0	5.8	4.2	85.8	4.2
	60	3.4	6.6	83.4	6.6	5.3	4.7	85.3	4.7	6.0	4.0	86.0	4.0
	70	3.6	6.4	83.6	6.4	5.5	4.5	85.5	4.5	6.3	3.7	86.3	3.7
	80	3.7	6.3	83.7	6.3	5.6	4.4	85.6	4.4	6.4	3.6	86.4	3.6
	90	3.9	6.1	83.9	6.1	5.8	4.2	85.8	4.2	6.6	3.4	86.6	3.4
	100	4.0	6.0	84.0	6.0	5.9	4.1	85.9	4.1	6.7	3.4	86.7	3.4
MNAR Inverse – Weak	*30*	2.3	7.7	82.6	7.4	3.8	6.2	83.8	6.2	4.7	5.3	84.7	5.3
	40	2.7	7.3	82.8	7.2	4.3	5.7	84.1	5.9	5.1	4.9	85.1	4.9
	50	3.0	7.0	83.0	7.0	4.6	5.4	84.6	5.4	5.6	4.4	85.6	4.4
	60	3.3	6.7	83.3	6.7	4.9	5.1	84.9	5.1	5.9	4.1	85.9	4.1
	70	3.5	6.5	83.5	6.5	5.3	4.7	85.3	4.7	6.1	3.9	86.1	3.9
	80	3.8	6.2	83.8	6.2	5.5	4.5	85.5	4.5	6.3	3.7	86.3	3.7
	90	4.0	6.0	84.0	6.0	5.6	4.4	85.6	4.4	6.5	3.5	86.5	3.5
	100	4.1	5.9	84.1	5.9	5.8	4.2	85.8	4.2	6.6	3.4	86.6	3.4
MNAR Inverse – Strong	*30*	2.2	7.8	82.6	7.4	3.7	6.3	83.7	6.3	4.5	5.5	84.5	5.5
	40	2.5	7.5	82.7	7.3	4.1	5.9	84.1	5.9	5.1	4.9	85.1	4.9
	50	2.8	7.2	82.8	7.2	4.5	5.5	84.5	5.5	5.5	4.5	85.5	4.5
	60	3.2	6.8	83.2	6.8	4.9	5.1	84.9	5.1	5.7	4.3	85.7	4.3
	70	3.4	6.6	83.4	6.6	5.1	4.9	85.1	4.9	6.0	4.0	86.0	4.0
	80	3.8	6.2	83.8	6.2	5.5	4.5	85.5	4.5	6.2	3.8	86.2	3.8
	90	3.9	6.1	83.9	6.1	5.6	4.4	85.6	4.4	6.5	3.5	86.5	3.5
	100	4.1	5.9	84.1	5.9	5.9	4.1	85.9	4.1	6.6	3.4	86.6	3.4

Abbreviations: MCAR = missing completely at random, MNAR = missing not at random, Doc. = documentation, TP = true positive, FP = false positive, TN = true negative, FN = false negative

Note: Sensitivity = TP/(TP+FN), Specificity = TN/(TN+FP), PVP = TP/(TP+FP), PVN = TN/(TN+FN)

Table A.2. Average proportion (%) of hospital high/low performer classification for top/bottom 20th percentile of rank-order (true positive, false positive, true negative, false negative) for different mechanisms of missing NIHSS data, stratified by hospital stroke volume (n=100, 300, and 500).

Mechanism of Missing NIHSS	***NIHSS Doc. Rate***	***Average Hospital High/Low Performer Classification (%)***											
		n=100				*n=300*				*n=500*			
		TP	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*
MCAR	*30*	19.1	20.9	39.4	20.6	23.4	16.6	43.4	16.6	25.9	14.1	45.9	14.1
	40	20.1	19.9	40.2	19.8	24.7	15.3	44.7	15.3	27.4	12.6	47.4	12.6
	50	20.9	19.1	41.1	18.9	25.8	14.2	45.8	14.2	28.3	11.7	48.3	11.7
	60	21.5	18.5	41.5	18.5	26.7	13.3	46.7	13.3	29.1	11.0	49.1	11.0
	70	22.2	17.8	42.2	17.8	27.3	12.7	47.3	12.7	29.8	10.2	49.8	10.2
	80	22.9	17.1	42.9	17.1	28.0	12.0	48.0	12.0	30.3	9.7	50.3	9.7
	90	23.5	16.5	43.5	16.5	28.6	11.4	48.6	11.4	30.9	9.1	50.9	9.1
	100	23.9	16.1	43.9	16.1	29.1	10.9	49.1	10.9	31.2	8.8	51.2	8.8
MNAR Direct – Weak	*30*	19.4	20.6	39.7	20.3	23.8	16.2	43.8	16.2	26.4	13.6	46.4	13.6
	40	20.5	19.5	40.6	19.4	25.0	15.0	45.0	15.0	27.6	12.4	47.6	12.4
	50	21.3	18.7	41.4	18.6	26.1	13.9	46.1	13.9	28.5	11.5	48.5	11.5
	60	21.9	18.1	41.9	18.1	26.9	13.1	46.9	13.1	29.4	10.6	49.4	10.6
	70	22.6	17.4	42.6	17.4	27.6	12.4	47.6	12.4	30.0	10.0	50.0	10.0
	80	23.1	16.9	43.1	16.9	28.1	11.9	48.1	11.9	30.5	9.5	50.5	9.5
	90	23.7	16.3	43.7	16.3	28.6	11.4	48.6	11.4	30.8	9.2	50.8	9.2
	100	24.1	15.9	44.1	15.9	29.0	11.0	49.0	11.0	31.2	8.8	51.2	8.8
MNAR Direct – Strong	*30*	18.1	21.9	39.8	20.2	22.6	17.4	42.6	17.4	25.0	15.0	45.0	15.0
	40	19.7	20.3	39.9	20.1	24.4	15.6	44.4	15.6	26.8	13.2	46.8	13.2
	50	21.6	18.4	41.6	18.4	26.5	13.5	46.5	13.5	29.0	11.0	49.0	11.0
	60	22.2	17.8	42.2	17.8	27.4	12.6	47.4	12.6	29.6	10.4	49.6	10.4
	70	22.8	17.2	42.8	17.2	27.8	12.2	47.8	12.2	30.1	9.9	50.1	9.9
	80	23.2	16.8	43.2	16.8	28.2	11.8	48.2	11.8	30.6	9.4	50.6	9.4
	90	23.7	16.3	43.7	16.3	28.6	11.4	48.6	11.4	30.9	9.1	50.9	9.1
	100	24.1	15.9	44.1	15.9	29.0	11.0	49.0	11.0	31.2	8.8	51.2	8.8
MNAR Inverse – Weak	*30*	18.6	21.4	39.8	20.2	23.3	16.7	43.3	16.7	25.5	14.5	45.5	14.5
	40	20.0	20.0	40.4	19.6	24.5	15.5	44.4	15.7	27.0	13.0	47.0	13.0
	50	20.7	19.3	40.7	19.3	25.5	14.5	45.5	14.5	28.1	11.9	48.1	11.9
	60	21.6	18.4	41.6	18.4	26.6	13.4	46.6	13.4	28.9	11.1	48.9	11.1
	70	22.3	17.7	42.3	17.7	27.3	12.7	47.3	12.7	29.6	10.4	49.6	10.4
	80	23.1	16.9	43.1	16.9	28.0	12.0	48.0	12.0	30.3	9.7	50.3	9.7
	90	23.4	16.6	43.4	16.6	28.5	11.5	48.5	11.5	30.8	9.2	50.8	9.2
	100	24.0	16.0	44.0	16.0	29.1	10.9	49.1	10.9	31.3	8.7	51.3	8.7
MNAR Inverse – Strong	*30*	17.8	22.2	39.7	20.3	22.7	17.3	42.7	17.3	24.9	15.1	44.9	15.1
	40	19.5	20.5	40.2	19.8	24.0	16.0	44.0	16.0	26.6	13.4	46.6	13.4
	50	20.5	19.5	40.5	19.5	25.2	14.8	45.2	14.8	27.7	12.3	47.7	12.3
	60	21.5	18.5	41.5	18.5	26.1	13.9	46.1	13.9	28.7	11.3	48.7	11.3
	70	22.0	18.0	42.0	18.0	27.0	13.0	47.0	13.0	29.5	10.5	49.5	10.5
	80	22.6	17.4	42.6	17.4	28.0	12.0	48.0	12.0	30.3	9.7	50.3	9.7
	90	23.5	16.5	43.5	16.5	28.6	11.4	48.6	11.4	30.6	9.4	50.6	9.4
	100	24.0	16.0	44.0	16.0	29.2	10.8	49.2	10.8	31.3	8.7	51.3	8.7

Abbreviations: MCAR = missing completely at random, MNAR = missing not at random, Doc. = documentation, TP = true positive, FP = false positive, TN = true negative, FN = false negative

Note: Sensitivity = TP/(TP+FN), Specificity = TN/(TN+FP), PVP = TP/(TP+FP), PVN = TN/(TN+FN)

Table A.3. Average absolute change in hospital RMSR Rankings (# of positions) in different scenarios of missing NIHSS data, stratified by quintile of true hospital ranking and hospital stroke volume (*n*=100, 300, and 500).

Mechanism of Missing NIHSS	***NIHSS Doc. Rate***	***Average Absolute Change Between True and Observed Rankings (# of Positions)***														
		n=100					*n=300*					*n=500*				
		1Q	*2Q*	*3Q*	*4Q*	*5Q*	*1Q*	*2Q*	*3Q*	*4Q*	*5Q*	*1Q*	*2Q*	*3Q*	*4Q*	*5Q*
MCAR	*30*	23.4	23.0	22.7	23.4	21.8	15.1	18.5	19.2	18.6	13.8	11.7	16.2	17.0	15.8	10.5
	40	21.3	21.7	22.0	22.0	19.5	13.2	17.3	18.0	17.3	12.0	10.2	14.8	15.7	14.4	9.0
	50	19.5	20.9	21.2	21.1	18.0	11.9	16.1	17.0	15.9	10.7	9.0	13.8	14.7	13.2	8.1
	60	18.1	20.4	21.1	20.6	16.5	10.6	15.2	16.2	14.9	9.7	8.2	12.9	13.8	12.4	7.3
	70	17.0	19.7	20.4	19.7	15.6	10.0	14.5	15.4	14.2	8.9	7.7	12.2	13.0	11.6	6.7
	80	16.2	18.9	19.9	19.2	14.4	9.3	13.7	14.7	13.6	8.3	7.1	11.5	12.4	10.9	6.2
	90	15.3	18.7	19.4	18.6	13.6	8.7	13.2	14.3	13.0	7.7	6.7	11.0	11.9	10.4	5.8
	100	14.6	18.1	19.0	18.0	13.0	8.3	12.7	13.7	12.4	7.3	6.3	10.5	11.5	10.0	5.5
MNAR Direct – Weak	*30*	22.4	22.9	22.3	22.6	21.3	14.4	18.1	19.0	18.2	13.2	11.3	15.6	16.5	15.5	10.2
	40	20.3	21.8	21.5	21.9	19.1	12.7	16.9	17.5	16.7	11.6	9.8	14.1	15.5	14.1	8.9
	50	18.7	20.8	21.1	20.9	17.3	11.3	16.0	16.6	15.5	10.2	8.9	13.3	14.2	13.1	7.9
	60	17.6	20.3	20.5	20.3	16.1	10.5	15.0	16.0	14.6	9.4	8.2	12.4	13.5	12.3	7.1
	70	16.5	19.8	20.0	19.7	15.1	9.8	14.4	15.3	14.1	8.7	7.7	11.8	12.9	11.5	6.6
	80	15.7	19.2	19.7	19.2	14.3	9.2	13.9	14.6	13.4	8.2	7.1	11.3	12.2	11.0	6.2
	90	14.9	18.7	19.1	18.5	13.6	8.6	13.2	14.1	12.8	7.6	6.7	10.8	11.7	10.5	5.8
	100	14.3	18.2	18.8	18.0	12.9	8.3	12.9	13.7	12.4	7.3	6.4	10.4	11.3	10.1	5.6
MNAR Direct – Strong	*30*	24.7	23.6	22.4	23.5	24.0	16.2	19.3	20.0	19.6	15.0	12.7	16.8	17.8	16.8	12.0
	40	21.6	22.4	22.0	22.4	20.6	13.6	17.6	18.3	17.7	12.4	10.7	15.0	16.2	15.0	9.8
	50	18.3	20.8	20.8	20.9	17.0	11.0	15.5	16.3	15.4	9.9	8.5	13.0	13.9	12.6	7.7
	60	17.1	20.1	20.3	20.1	15.7	10.2	14.7	15.6	14.4	9.1	7.9	12.2	13.2	12.0	6.9
	70	16.3	19.5	20.0	19.4	15.0	9.5	14.2	14.9	13.7	8.5	7.3	11.6	12.7	11.4	6.5
	80	15.5	19.0	19.4	18.8	14.2	9.0	13.6	14.6	13.3	8.1	7.0	11.1	12.2	10.8	6.1
	90	14.8	18.6	19.0	18.5	13.5	8.6	13.3	14.1	12.9	7.6	6.7	10.7	11.7	10.4	5.8
	100	14.3	18.2	18.8	18.0	12.9	8.3	12.9	13.7	12.4	7.3	6.4	10.4	11.3	10.1	5.6
MNAR Indirect – Weak	*30*	23.9	23.1	22.4	23.5	22.6	15.4	18.7	19.5	18.9	14.2	12.4	16.3	17.3	16.5	11.0
	40	21.4	21.8	21.7	22.1	20.2	13.7	17.4	18.4	17.6	12.2	10.4	15.2	16.1	14.9	9.6
	50	19.7	21.1	21.6	21.5	18.2	12.2	16.6	17.1	16.2	11.0	9.3	14.0	15.0	13.7	8.3
	60	18.1	20.7	20.8	20.7	16.7	11.2	15.5	16.1	15.2	9.9	8.4	13.1	14.0	12.7	7.5
	70	17.0	19.8	20.2	20.1	15.5	10.3	14.7	15.5	14.5	9.0	7.8	12.2	13.2	11.7	6.9
	80	15.8	19.2	19.7	19.2	14.5	9.4	13.9	14.9	13.7	8.4	7.1	11.5	12.5	11.2	6.3

Table A.3. (cont'd) Average absolute change in hospital RMSR Rankings (# of positions) in different scenarios of missing NIHSS data, stratified by quintile of true hospital ranking and hospital stroke volume (*n*=100, 300, and 500).

Mechanism of Missing NIHSS	***NIHSS Doc. Rate***	***Average Absolute Change Between True and Observed Rankings (# of Positions)***														
		n=100						*n=300*						*n=500*		
		1Q	*2Q*	*3Q*	*1Q*	*2Q*	*3Q*	*1Q*	*2Q*	*3Q*	*1Q*	*2Q*	*3Q*	*1Q*	*2Q*	*3Q*
MNAR	*90*	15.1	18.7	19.1	18.7	13.7	8.9	13.3	14.2	13.1	7.8	6.7	11.0	12.0	10.6	5.9
Indirect – Weak	*100*	14.3	18.1	18.6	18.1	12.9	8.3	12.8	13.7	12.6	7.4	6.3	10.5	11.4	10.1	5.6
	30	24.7	23.3	22.3	23.6	23.6	16.2	19.4	19.7	19.4	14.9	12.8	17.0	18.0	16.8	11.7
	40	22.3	22.4	22.0	22.7	20.9	14.3	18.0	18.6	17.8	12.9	11.1	15.6	16.6	15.3	9.9
	50	20.3	21.8	21.8	21.8	18.8	12.5	16.8	17.7	16.7	11.5	9.9	14.4	15.4	14.1	8.6
MNAR	*60*	18.7	20.7	21.0	20.7	17.1	11.6	15.6	16.8	15.6	10.2	8.9	13.4	14.3	13.0	7.8
Indirect – Strong	*70*	17.4	20.2	20.3	20.2	15.9	10.5	14.8	16.0	14.6	9.3	7.9	12.4	13.6	12.1	7.0
	80	16.4	19.4	19.8	19.4	14.8	9.7	14.0	15.1	13.8	8.5	7.2	11.7	12.7	11.4	6.4
	90	15.3	18.6	19.2	18.8	13.7	8.9	13.4	14.4	13.1	7.9	6.8	11.1	12.1	10.8	6.0
	100	14.3	18.1	18.6	18.1	12.9	8.2	12.7	13.7	12.4	7.4	6.3	10.5	11.4	10.1	5.6

Quintiles: 1Q:1-20, 2Q: 21-40, 3Q: 41-60, 4Q: 61-80, 5Q: 81-100

Appendix B: Supplementary Figures

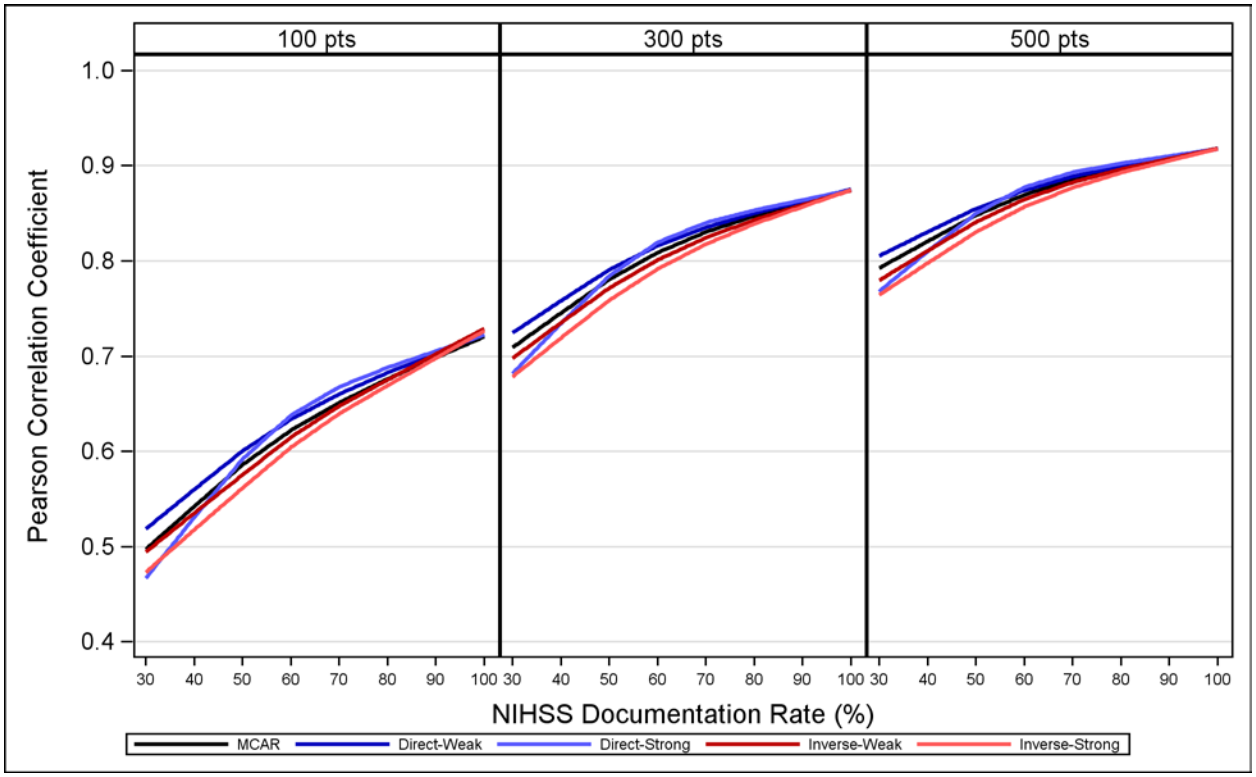

Figure B.1. Pearson correlation coefficients between true rankings and RSMR rankings as NIHSS documentation increases under different mechanisms of missing NIHSS data. Results are stratified by hospital stroke volume.

Appendix C: IRB Determination

MICHIGAN STATE
UNIVERSITY

Determined Not "Human Subjects"

November 7, 2014

Michael Thompson
Email: mthompson@epi.msu.edu

Re: Determination of Human Subject Research
Assessing the Impact of Missing Data on Hospital Performance Profiling

Dear Michael Thompson:

It has been determined that the activity described in your application to the IRB submitted November 3, 2014 does not meet the definition of "human subjects" as defined by the U.S. Department of Health and Human Services (DHHS) regulations for the protection of human research subjects.

Human Subject
For DHHS, "human subject" means "a living individual about whom an investigator (whether professional or student) conducting research obtains: (1) Data through intervention or interaction with the individual, or (2) Identifiable private information." [45 CFR 46.102(f)].

After reviewing the information you have provided, it has been determined that:
☒ Living individuals are not involved
☒ The activity is not "about" the living individual
☒ Will not obtain data through interaction or intervention or private identifiable information

You project involves the analysis of the Michigan Stroke Registry, which has been de-identified by MDCH as well as the use of computer simulation models neither of which meet the definition of human subjects.

Hence, your activity does not involve human subjects.

Therefore, the federal regulations for the protection of human subjects would not apply to your project and you do not need MSU IRB approval to proceed. However, please note that while MSU IRB approval is not required, other federal, state, or local regulations or requirements or ethical or professional standards may still be applicable based on your activity.

If any of these circumstances change, please contact the IRB as your activity may involve human subject research and require IRB approval.

If you have any further questions, please contact the MSU IRB office at 517-355-2180.

Sincerely,

Ashir Kumar, M.D.
Chair, Biomedical and Health IRB

Office of Regulatory Affairs
Human Research Protection Programs

Biomedical & Health Institutional Review Board (BIRB)

Community Research Institutional Review Board (CRIRB)

Social Science Behavioral/Education Institutional Review Board (SIRB)

Olds Hall
408 West Circle Drive
Room 207
East Lansing, MI 48824
(517) 355-2180
Fax: (517) 432-4503
Email: irb@msu.edu

V14.1

Appendix D: Example Data Generation SAS Code

```
/*************************************************************************
Title:      Data Generation for Simulation Modeling
Date: 11/10/14
Descr.: SAS code to generate data that is similar in structure to the
Michigan Stroke Registry (MSR).  Does not include changes in Risk score
distribution as noted by the primary stroke center status.  Does include
differences in missing NIHSS frequency
*************************************************************************/
/* Suppress log - nonotes=no log statements, notes=log statements */
options nonotes;

libname sim "L:\MASCOTS\Mike\Dissertation\Analysis\Simulation Runs";

/* Set # of samples to run (S) and Hospitals (M) per sample */
%Let S=500;
%Let M=100;

data init;
call streaminit(02052015);
      do Sampleid = 1 to &S ;
      /* Number of Patients per Hospital */
      do hospid = 1 to &M;
      /* b0 --> Assigned hospital random intercept "true ranking" */
      b0 = rand("Normal", 0, sqrt(0.13));
      do vol = 500;
            hospSRS = rand("Normal", 0, sqrt(1.5));
            do rep = 1 to vol;
                  /* SRS = Sub-risk score */
                  ptSRS = rand("Normal", 0, sqrt(68.0));
                  muSRS = ptsrs + hospSRS;
                  SRS = 21.4 + muSRS;
                  output;
            end;
      end;
      end;
      end;
run;

data init;
      set init;
      PatID = _N_;
      if SRS<0 then delete; if SRS>44 then delete;

      /* Generate NIHSS Categories and RS weights from Eta - based on ordinal
      model cut points. Note:  To change frequency of categories, adjust cut
      points as necessary */
```

```
      /* NIHSS */
      eps = rand("Normal", 0, 1);
      eta = -0.050*SRS + eps;
      if eta>=-0.63 then nih=1;           if nih=1 then rsnih=0;
      if -0.63> eta >=-1.26 then nih=2;   if nih=2 then rsnih=10;
      if -1.26> eta >=-1.79 then nih=3;   if nih=3 then rsnih=21;
      if -1.79> eta >=-2.16 then nih=4;   if nih=4 then rsnih=37;
      if -2.16> eta >=-2.55 then nih=5;   if nih=5 then rsnih=48;
      if -2.55> eta >=-2.98 then nih=6;   if nih=6 then rsnih=56;
      if eta<-2.98 then nih=7;            if nih=7 then rsnih=65;

      /* Total Risk Score (TRS) and probability from algorithm calculated */
      TRS = SRS + rsNIH;
      logitphat = -4.4 + 0.054*trs + b0;
      phat = exp(logitphat) / (1 + exp(logitphat)  );

      /* Calc of Patient Mortality - use parameters from registry model*/
      died = rand("Bernoulli", phat);

      drop hospsrs ptsrs musrs eta eps;
      do doc=30 to 100 by 10;
            output;
      end;
run;

data Miss;
      set init;

      /* Missingness Scenario */

      /* MCAR */
      /*if doc=100 then obs=rand("Bernoulli", 1.00);
      if doc=90 then obs=rand("Bernoulli", 0.90);
      if doc=80 then obs=rand("Bernoulli", 0.80);
      if doc=70 then obs=rand("Bernoulli", 0.70);
      if doc=60 then obs=rand("Bernoulli", 0.60);
      if doc=50 then obs=rand("Bernoulli", 0.50);
      if doc=40 then obs=rand("Bernoulli", 0.40);
      if doc=30 then obs=rand("Bernoulli", 0.30);*/


if doc=100 then obslow=1;
if doc=90 then obslow=rand("Bernoulli", 1/(1+exp(-(2.00 + .095*nih))));
if doc=80 then obslow=rand("Bernoulli", 1/(1+exp(-(1.15 + .095*nih))));
if doc=70 then obslow=rand("Bernoulli", 1/(1+exp(-(0.60 + .095*nih))));
if doc=60 then obslow=rand("Bernoulli", 1/(1+exp(-(0.17 + .095*nih))));
if doc=50 then obslow=rand("Bernoulli", 1/(1+exp(-(-0.25 + 0.095*nih))));
if doc=40 then obslow=rand("Bernoulli", 1/(1+exp(-(-0.65 + 0.095*nih))));
if doc=30 then obslow=rand("Bernoulli", 1/(1+exp(-(-1.10 + 0.095*nih))));
if doc=100 then obshigh=1;
```

```
if doc=90 then obshigh=rand("Bernoulli", 1/(1+exp(-(1.65 + 0.225*nih))));
if doc=80 then obshigh=rand("Bernoulli", 1/(1+exp(-(0.85 + 0.225*nih))));
if doc=70 then obshigh=rand("Bernoulli", 1/(1+exp(-(0.29 + 0.225*nih))));
if doc=60 then obshigh=rand("Bernoulli", 1/(1+exp(-(-0.15 + 0.225*nih))));
if doc=50 then obshigh=rand("Bernoulli", 1/(1+exp(-(-0.58 + 0.225*nih))));
if doc=40 then obshigh=rand("Bernoulli", 1/(1+exp(-(-1.45 + 0.225*nih))));
if doc=30 then obshigh=rand("Bernoulli", 1/(1+exp(-(-2.00 + 0.225*nih))));

run;
proc sort data=miss; by doc sampleid; run;


* Step 0: Import data set.  The dataset "original" will be used throughout
the analysis as necessary, no need to change the name of the variable.  Also,
create hospid variable based on whatever hospital id is being used in your
data.  This dataset uses the outcome of "died" as binary a outcome (died=1,
alive=0).  Change outcome as necessary, but be sure that event=1 for the
analysis.  This analysis uses a single risk score variable, x1, as the
independent predictor, but more can be inserted as needed.;

/*******P/E Method - Hierarchical Logist Regression Model ***************/
/* LOW */
* Step 1:  Calculate predicted probability of mortality.  NOTE: This
probability includes hospital random effect in the calculation. (i.e. blup
statement);

ods trace on;
ods select solutionr /*parameterestimates*/ ;
title "Model for P/E Ratio Rankings";
proc glimmix data=miss initglm  ;
      where obslow=1;
      by doc SampleID;
      class hospid;
      model died(event='1') = TRS / dist=binary link=logit ddfm=bw ;
      random int / subject=hospid s;
      nloptions tech=nrridg;
      output out=gmxout1 pred(blup ilink)=phatpred pred(noblup ilink)=phatexp
;
      ods output "Solution for Random Effects"=solutionr
            /*"Solutions for Fixed Effects"=param*/;
run;

data solutionr;
      set solutionr;
      newvar=compress(subject,'hospid ');
      hospid=newvar*1;
      drop newvar subject effect;;
run;
```

```
* Step 2:  Calculate predicted, expected and observed deaths for each
hospital. This is done by summing the predicted probability, expected
probability, and observed deaths for each patient in the hospital.;

proc means data=gmxout1 noprint;
      by doc SampleID hospid;
      output out=PE sum(phatpred)=pred sum(phatexp)=exp sum(died)=obsdied
            mean(b0)=b0 ;
run;

* Step 3:  Calculate: PMR/EMR, P/E ratio, SMR-P/E, OMR/EMR, O/E ratio, SMR-
O/E, observed hospital random intercept and outlier status;
data PElow;
      merge PE solutionr;
      by doc sampleid hospid;
      pmr=pred/_FREQ_*100;
      emr=exp/_FREQ_*100;
      PEratio=pmr/emr;
      SMRPE=15*peratio;
      omr=obsdied/_freq_*100;
      OEratio=omr/emr;
      SMROE=15*oeratio;
      hosp=put(hospid, 3.);
      obsb0=estimate;
      drop _type_ df estimate;
run;

/* Rankings of Hospitals */
title "True rankings";
proc rank data=pelow out=pelowrank;
      by doc sampleid;
      var b0;
      ranks b0rank;
run;

title "Observed RI Hospital rankings";
proc rank data=pelowrank out=pelowrank;
      by doc sampleid;
      var obsb0;
      ranks obsb0rank;
run;



title "RSMR - P/E rankings";
proc rank data=pelowrank out=pelowrank;
      by doc sampleid;
      var SMRPE;
      ranks SMRPErank;
run;
```

```
title "RSMR - O/E rankings";
proc rank data=pelowrank out=pelowrank;
      by doc sampleid;
      var SMROE;
      ranks SMROErank;
run;
```

Appendix E: Example Simulation Assessment SAS Code

```
/********* Assessment of Observed v True Random Intercepts ***************/

/* Se, Sp, PVP, PVN for True Outlier status */
proc means data=pemain2 noprint;
      class size mechanism doc sampleid;
      output out=means3 sum(tp5rank)=tp5rank sum(tn5rank)=tn5rank
                  sum(fn5rank)=fn5rank sum(fp5rank)=fp5rank
                  sum(tp20rank)=tp20rank sum(tn20rank)=tn20rank
                  sum(fn20rank)=fn20rank sum(fp20rank)=fp20rank
                  sum(pe20)=pe20 sum(pe5)=pe5;
run;
data means4;
      set means3;
      if _type_ ne 15 then delete;
      Se5=(tp5rank/(tp5rank+fn5rank))*100;
      Se20=(tp20rank/(tp20rank+fn20rank))*100;
      Sp5=(tn5rank/(tn5rank+fp5rank))*100;
      Sp20=(tn20rank/(tn20rank+fp20rank))*100;
      PVP5=(tp5rank/(tp5rank+fp5rank))*100;
      PVP20=(tp20rank/(tp20rank+fp20rank))*100;
      PVN5=(tn5rank/(tn5rank+fn5rank))*100;
      PVN20=(tn20rank/(tn20rank+fn20rank))*100;
      pe20=pe20/_freq_*100;
      pe5=pe5/_freq_*100;
run;
proc means data=means4 noprint;
      class size mechanism doc;
      output out=acc2   mean(Se5)=Se5 stderr(Se5)=Se5Err
                              mean(Se20)=Se20 stderr(Se20)=Se20Err
                              mean(Sp5)=Sp5 stderr(Sp5)=Sp5Err
                              mean(Sp20)=Sp20 stderr(Sp20)=Sp20Err
                              mean(PVP5)=PVP5 stderr(PVP5)=PVP5Err
                              mean(PVP20)=PVP20 stderr(pvp20)=PVP20Err
                              mean(PVN5)=PVN5 stderr(PVN5)=PVN5Err
                              mean(PVN20)=PVN20 stderr(PVN20)=PVN20Err
                              mean(pe20)=pe20 stderr(pe20)=pe20err
                              mean(pe5)=pe5 stderr(pe5)=pe5err
                              mean(tp5rank)=TP5 mean(tn5rank)=TN5
                              mean(fp5rank)=FP5 mean(fn5rank)=FN5
                              mean(tp20rank)=TP20 mean(tn20rank)=TN20
                              mean(fp20rank)=FP20 mean(fn20rank)=FN20;
run;
data accuracy;
      set acc2;
      if _type_ ne 7 then delete;
run;
```

```
/* Sensitivity */
ods graphics / imagefmt=png height=4in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=assessment2 noautolegend;
      panelby size/ columns=3 novarname;
      loess x=doc y=se5_1 / nomarkers legendlabel="MCAR 5%"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=se5_2 / nomarkers legendlabel="Direct-Weak 5%"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=se5_3 / nomarkers legendlabel="Direct-Strong 5%"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=se5_4 / nomarkers legendlabel="Inverse-Weak 5%"
lineattrs=(thickness=2 color=dark_red pattern=1);
      loess x=doc y=se5_5 / nomarkers legendlabel="Inverse-Strong 5%"
lineattrs=(thickness=2 color=light_red pattern=1);
      loess x=doc y=se20_1 / nomarkers legendlabel="MCAR 20%"
lineattrs=(thickness=2 color=black pattern=4);
      loess x=doc y=se20_2 / nomarkers legendlabel="Direct-Weak 20%"
lineattrs=(thickness=2 color=dark_blue pattern=4);
      loess x=doc y=se20_3 / nomarkers legendlabel="Direct-Strong 20%"
lineattrs=(thickness=2 color=light_blue pattern=4);
      loess x=doc y=se20_4 / nomarkers legendlabel="Inverse-Weak 20%"
lineattrs=(thickness=2 color=dark_red pattern=4);
      loess x=doc y=se20_5 / nomarkers legendlabel="Inverse-Strong 20%"
lineattrs=(thickness=2 color=light_red pattern=4);
      rowaxis label="Sensitivity (%)" values=(20 to 80 by 5) grid;
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom across=5 down=2 valueattrs=(size=6) ;
      format size size.;
run;

/* Specificity */

ods graphics / imagefmt=png height=4in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=assessment2 noautolegend;
      panelby size/ columns=3 novarname;
      loess x=doc y=sp5_1 / nomarkers legendlabel="MCAR 5%"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=sp5_2 / nomarkers legendlabel="Direct-Weak 5%"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=sp5_3 / nomarkers legendlabel="Direct-Strong 5%"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=sp5_4 / nomarkers legendlabel="Inverse-Weak 5%"
lineattrs=(thickness=2 color=dark_red pattern=1);
```

```
      loess x=doc y=sp5_5 / nomarkers legendlabel="Inverse-Strong 5%"
lineattrs=(thickness=2 color=light_red pattern=1);
      loess x=doc y=sp20_1 / nomarkers legendlabel="MCAR 20%"
lineattrs=(thickness=2 color=black pattern=4);
      loess x=doc y=sp20_2 / nomarkers legendlabel="Direct-Weak 20%"
lineattrs=(thickness=2 color=dark_blue pattern=4);
      loess x=doc y=sp20_3 / nomarkers legendlabel="Direct-Strong 20%"
lineattrs=(thickness=2 color=light_blue pattern=4);
      loess x=doc y=sp20_4 / nomarkers legendlabel="Inverse-Weak 20%"
lineattrs=(thickness=2 color=dark_red pattern=4);
      loess x=doc y=sp20_5 / nomarkers legendlabel="Inverse-Strong 20%"
lineattrs=(thickness=2 color=light_red pattern=4);
      rowaxis label="Specificity (%)" values=(60 to 100 by 5) grid;
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom across=5 down=2 valueattrs=(size=6) ;
      format size size.;
run;

/* PVP */

ods graphics / imagefmt=png height=4in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=assessment2 noautolegend;
      panelby size/ columns=3 novarname;
      loess x=doc y=pvp5_1 / nomarkers legendlabel="MCAR 5%"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=pvp5_2 / nomarkers legendlabel="Direct-Weak 5%"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=pvp5_3 / nomarkers legendlabel="Direct-Strong 5%"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=pvp5_4 / nomarkers legendlabel="Inverse-Weak 5%"
lineattrs=(thickness=2 color=dark_red pattern=1);
      loess x=doc y=pvp5_5 / nomarkers legendlabel="Inverse-Strong 5%"
lineattrs=(thickness=2 color=light_red pattern=1);
      loess x=doc y=pvp20_1 / nomarkers legendlabel="MCAR 20%"
lineattrs=(thickness=2 color=black pattern=4);
      loess x=doc y=pvp20_2 / nomarkers legendlabel="Direct-Weak 20%"
lineattrs=(thickness=2 color=dark_blue pattern=4);
      loess x=doc y=pvp20_3 / nomarkers legendlabel="Direct-Strong 20%"
lineattrs=(thickness=2 color=light_blue pattern=4);
      loess x=doc y=pvp20_4 / nomarkers legendlabel="Inverse-Weak 20%"
lineattrs=(thickness=2 color=dark_red pattern=4);
      loess x=doc y=pvp20_5 / nomarkers legendlabel="Inverse-Strong 20%"
lineattrs=(thickness=2 color=light_red pattern=4);
      rowaxis label="Predictive Value Positive (%)" values=(20 to 80 by 5)
grid;
```

```
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom across=5 down=2 valueattrs=(size=6) ;
      format size size.;
run;

/* PVN */

ods graphics / imagefmt=png height=4in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=assessment2 noautolegend;
      panelby size/ columns=3 novarname;
      loess x=doc y=pvn5_1 / nomarkers legendlabel="MCAR 5%"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=pvn5_2 / nomarkers legendlabel="Direct-Weak 5%"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=pvn5_3 / nomarkers legendlabel="Direct-Strong 5%"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=pvn5_4 / nomarkers legendlabel="Inverse-Weak 5%"
lineattrs=(thickness=2 color=dark_red pattern=1);
      loess x=doc y=pvn5_5 / nomarkers legendlabel="Inverse-Strong 5%"
lineattrs=(thickness=2 color=light_red pattern=1);
      loess x=doc y=pvn20_1 / nomarkers legendlabel="MCAR 20%"
lineattrs=(thickness=2 color=black pattern=4);
      loess x=doc y=pvn20_2 / nomarkers legendlabel="Direct-Weak 20%"
lineattrs=(thickness=2 color=dark_blue pattern=4);
      loess x=doc y=pvn20_3 / nomarkers legendlabel="Direct-Strong 20%"
lineattrs=(thickness=2 color=light_blue pattern=4);
      loess x=doc y=pvn20_4 / nomarkers legendlabel="Inverse-Weak 20%"
lineattrs=(thickness=2 color=dark_red pattern=4);
      loess x=doc y=pvn20_5 / nomarkers legendlabel="Inverse-Strong 20%"
lineattrs=(thickness=2 color=light_red pattern=4);
      rowaxis label="Predictive Value Negative (%)" values=(60 to 100 by 5)
grid;
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom across=5 down=2 valueattrs=(size=6) ;
      format size size.;
run;
```

```
/* Spearman Correlation */
ods graphics / imagefmt=png height=4in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=assessment2 noautolegend;
      panelby size/ columns=3 novarname;
      loess x=doc y=spcorr1 / nomarkers legendlabel="MCAR"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=spcorr2 / nomarkers legendlabel="Direct-Weak"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=spcorr3 / nomarkers legendlabel="Direct-Strong"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=spcorr4 / nomarkers legendlabel="Inverse-Weak"
lineattrs=(thickness=2 color=dark_red pattern=1);
      loess x=doc y=spcorr5 / nomarkers legendlabel="Inverse-Strong"
lineattrs=(thickness=2 color=light_red pattern=1);
      rowaxis label="Spearman Rank Correlation Coefficient" values=(0.4 to 1
by 0.1) grid;
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom across=5 down=1 valueattrs=(size=6) ;
      format size size.;
run;
/* Pearson Correlation */
ods graphics / imagefmt=png height=4in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=assessment2 noautolegend;
      panelby size/ columns=3 novarname;
      loess x=doc y=pearcorr1 / nomarkers legendlabel="MCAR"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=pearcorr2 / nomarkers legendlabel="Direct-Weak"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=pearcorr3 / nomarkers legendlabel="Direct-Strong"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=pearcorr4 / nomarkers legendlabel="Inverse-Weak"
lineattrs=(thickness=2 color=dark_red pattern=1);
      loess x=doc y=pearcorr5 / nomarkers legendlabel="Inverse-Strong"
lineattrs=(thickness=2 color=light_red pattern=1);
      rowaxis label="Pearson Correlation Coefficient" values=(0.4 to 1 by
0.1) grid;
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom across=5 down=1 valueattrs=(size=6) ;
      format size size.;
run;
```

```
/*********** Difference between True/Observed Rankings *********/

ods graphics / imagefmt=png height=7.5in width=6.5in antialias=on
antialiasmax=1000;
ods listing device=png image_dpi=300;
title font=arial;
proc sgpanel data=difftable2 noautolegend;
      panelby size rankcat  / colheaderpos=top layout=lattice onepanel
novarname ;
      loess x=doc y=pediff1 / nomarkers legendlabel="MCAR"
lineattrs=(thickness=2 color=black pattern=1);
      loess x=doc y=pediff2 / nomarkers legendlabel="Direct-Weak"
lineattrs=(thickness=2 color=dark_blue pattern=1);
      loess x=doc y=pediff3 / nomarkers legendlabel="Direct-Strong"
lineattrs=(thickness=2 color=light_blue pattern=1);
      loess x=doc y=pediff4 / nomarkers legendlabel="Inverse-Weak"
lineattrs=(thickness=2 color=dark_red pattern=1);
      loess x=doc y=pediff5 / nomarkers legendlabel="Inverse-Strong"
lineattrs=(thickness=2 color=light_red pattern=1);
      rowaxis grid label="Absolute Change in Hospital RMSR Rankings (# of
Positions)"
            values=(4 to 26 by 4) valueattrs=(size=6);
      colaxis label="NIHSS Documentation Rate (%)" values=(30 to 100 by 10)
valueattrs=(size=6);
      keylegend / position=bottom valueattrs=(size=6)  ;
      format size size. mechanism mechanism. rankcat rankcat.;
run;
```

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Go AS, Mozaffarian D, Roger VL, et al. Heart Disease and Stroke Statistics—2013 Update: A Report From the American Heart Association. *Circulation.* January 1, 2013 2013;127(1):e6-e245.

2. AHRQ. Household component summary table. Table 4: Total Expenses and Percent Distribution for Selected Conditions by Source of Payment: United States, 2011. http://meps.ahrq.gov/mepsweb/data_stats/tables_compendia_hh_interactive.jsp?_SERVICE=MEPSSocket0&_PROGRAM=MEPSPGM.TC.SAS&File=HCFY2011&Table=HCFY2011_CNDXP_D&_Debug=. Accessed November, 15, 2013.

3. Stepanova M, Venkatesan C, Altaweel L, Mishra A, Younossi ZM. Recent trends in inpatient mortality and resource utilization for patients with stroke in the United States: 2005-2009. *Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association.* 2013;22:491-499.

4. Wier LM, Andrews RM. *The National Hospital Bill: The Most Expensive Conditions by Payer, 2008.* Rockville, MD: Agency for Healthcare Research and Quality;2011.

5. Centers for Medicare & Medicaid Services. Hospital Inpatient Quality Reporting Program. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html. Accessed January 8, 2014, 2014.

6. Centers for Medicare & Medicaid Services. Hospital Value-Based Purchasing. 2013; http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/index.html?redirect=/hospital-value-based-purchasing. Accessed January 8, 2014, 2014.

7. O'Kane ME. Performance-Based Measures: The Early Results Are In. *Journal of Managed Care Pharmacy.* 2007;13(2(Suppl S-b)):S3-S6.

8. Moses III H, Matheson DM, Dorsey E, George BP, Sadoff D, Yoshimura S. The Anatomy of Health Care in the United States. *JAMA.* 2013;310(18):1947-1964.

9. Sisko AM, Keehan SP, Cuckler GA, et al. National health expenditure projections, 2013-23: faster growth expected with expanded coverage and improving economy. *Health Aff. (Millwood).* Oct 1 2014;33(10):1841-1850.

10. James J. Health Policy Brief: Pay-for-Performance. 2012; http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=78. Accessed November 11, 2014, 2014.

11. Centers for Medicare & Medicaid Services. Outcome Measures. http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html. Accessed October 24, 2014, 2014.

12. Centers for Medicare & Medicaid Services. What is Hospital Compare? *Hospital Compare* 2013; http://www.medicare.gov/hospitalcompare/About/What-Is-HOS.html. Accessed January 14, 2014.

13. QualityNet. Measure Comparison (Inpatient Hospital Quality Measures). 2014; https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1138900298473. Accessed November 19, 2014.

14. QualityNet. Measures: Hospital Value-Based Purchasing. 2014; http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228772237361. Accessed December 12, 2014.

15. Centers for Medicare & Medicaid Services. Outcome Measures. *Hospital Quality Initiative* 2014; http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html. Accessed November 11, 2014, 2014.

16. Spivack SB, Bernheim SM, Forman HP, Drye EE, Krumholz HM. Hospital cardiovascular outcome measures in federal pay-for-reporting and pay-for-performance programs: a brief overview of current efforts. *Circ Cardiovasc Qual Outcomes.* Sep 2014;7(5):627-633.

17. Centers for Medicare & Medicaid Services. Readmissions Reduction Program. 2014; http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html. Accessed December 12, 2014, 2014.

18. Dorsey K, Grady JN, Wang Y, et al. *2014 Measures Updates and Specifications Report: Hospital-Level 30-Day Risk Standardized Mortality Measures.* Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE);2014.

19. Ash AS, Fienberg SE, Louis TA, Normand S-LT, Stukel TA, Utts J. *Statistical Issues in Assessing Hospital Performance.* COPSS-CMS White Paper Committee;2012.

20. Normand S-LT, Glickman ME, Gatsonis CA. Statistical Methods for Profiling Providers of Medical Care : Issues and Applications. *Journal of the American Statistical Association.* 1997;92:803-814.

21. Iezzoni LI. *Risk Adjustment for Measuring Health Care Outcomes.* Third Edition ed. Chicago, IL: Health Administration Press; 2003.

22. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology an. *Circulation.* 2006;113:456-462.

23. Krumholz HM, Wang Y, Mattera Ja, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. *Circulation.* 2006;113:1693-1701.

24. Krumholz HM, Wang Y, Mattera Ja, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation.* 2006;113:1683-1692.

25. Harrell FE, Lee KL, Mark DB. Multivariable progostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 1996;15(4):361-387.

26. Katzan IL, Spertus J, Bettger JP, et al. Risk Adjustment of Ischemic Stroke Outcomes for Comparing Hospital Performance: A Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke.* January 23, 2014 2014.

27. Daniels MJ, Gatsonis CA. Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization. *Journal of the American Statistical Association.* 1999;94(445):29-42.

28. Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. *Statistical Science.* 2007;22:206-226.

29. Austin PC, Alter Da, Tu JV. The use of fixed- and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. *Med. Decis. Making.* 2003;23:526-539.

30. Christiansen CL, Morris CN. Improving the Statistical Approach to Health Care Provider Profiling. *Ann. Intern. Med.* 1997;127(8_Part_2):764-768.

31. Goldstein H, Spiegelhalter DJ. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society.* 1996;159(3):385-443.

32. Shahian DM, Torchiana DF, Shemin RJ, Rawn JD, Normand S-LT. Massachusetts Cardiac Surgery Report Card: Implications of Statistical Methodology. *The Annals of Thoracic Surgery.* 12// 2005;80(6):2106-2113.

33. Jones HE, Spiegelhalter DJ. The Identification of "Unusual" Health-Care Providers From a Hierarchical Model. *The American Statistician.* 2011;65(3):154-163.

34. Ieva F, Paganoni AM. Detecting and visualizing outliers in provider profiling via funnel plots and mixed effect models. *Health Care Manag. Sci.* Jan 10 2014.

35. Shahian DM, He X, Jacobs JP, et al. Issues in quality measurement: target population, risk adjustment, and ratings. *Ann. Thorac. Surg.* Aug 2013;96(2):718-726.

36. Shahian DM, Iezzoni LI, Meyer GS, Kirle L, Normand S-LT. Hospital-wide Mortality as a Quality Metric: Conceptual and Methodological Challenges. *Am. J. Med. Qual.* March 1, 2012 2012;27(2):112-123.

37. McCrum ML, Joynt KE, Orav EJ, Gawande AA, Jha AK. Mortality for publicly reported conditions and overall hospital mortality rates. *JAMA internal medicine.* Jul 22 2013;173(14):1351-1357.

38. Jha AK, Li Z, Orav EJ, Epstein AM. Care in U.S. hospitals--the Hospital Quality Alliance program. *The New England journal of medicine.* 2005;353:265-274.

39. Mackenzie SJ, Goldmann DA, Perla RJ, Parry GJ. Measuring Hospital-Wide Mortality - Pitfalls and Potential. *J. Healthc. Qual.* 2014;00(0):0.

40. Mohammed MA, Deeks JJ, Girling A, et al. Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. *BMJ.* 2009-03-18 15:01:48 2009;338.

41. Shahian DM, Wolf RE, Iezzoni LI, Kirle L, Normand S-LT. Variability in the measurement of hospital-wide mortality rates. *The New England journal of medicine.* 2010;363:2530-2539.

42. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JP, Mackiernan Y. Judging Hospitals by Severity-Adjusted Mortality Rates: The Influence of the Severity-Adjustment Method. *Am. J. Public Health.* 1996;86(10):1379-1387.

43. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, Mackiernan Y. Severity Measurement Methods and Judging Hospital Death Rates for Pneumonia. *Med. Care.* 1996;34(1):11-28.

44. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Qual Saf.* Sep 2012;21(9):737-745.

45. Guru V, Tu JV, Etchells E, et al. Relationship Between Preventability of Death After Coronary Artery Bypass Graft Surgery and All-Cause Risk-Adjusted Mortality Rates. *Circulation.* 2008;117:2969-2976.

46. Centers for Medicare & Medicaid Services (CMS). Medicare Program; Hospital Prospective Payment System and Fiscal Year 2014 Rates. *Fed. Regist.* August 2, 2013 2013;78(160):50495-51040.

47. Arnett DK. Letter to Centers for Medicare & Medicaid Services. *Re: Docket No. CMS-1599-P.* 2013. https://www.heart.org/idc/groups/heart-public/@wcm/@adv/documents/downloadable/ucm_453664.pdf. Accessed June 25, 2013.

48. Fonarow GC, Alberts MJ, Broderick JP, et al. Stroke outcomes measures must be appropriately risk adjusted to ensure quality care of patients. *Stroke.* 2014;45(5):1589-1601.

49. Centers for Medicare & Medicaid Services. Quality Data Reporting Requirements for Specific Providers and Suppliers; Final Rule. In: Department of Health and Human Services, ed. Vol 78: Federal Register; 2013:50774-50906.

50. National Stroke Association. NIH Stroke Scale. 2014; http://www.stroke.org/site/PageServer?pagename=NIHSS.

51. Kasner SE. Clinical interpretation and use of stroke scales. *The Lancet Neurology.* 2006;5(7):603-612.

52. Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PLoS ONE.* 2011;6(4):e17401.

53. Fonarow GC, Pan W, Saver JL, et al. Comparison of 30-day mortality models for profiling hospital performance in acute ischemic stroke with vs without adjustment for stroke severity. *JAMA.* 2012;308(3):257-264.

54. Smith EE, Shobha N, Dai D, et al. Risk score for in-hospital ischemic stroke mortality derived and validated within the Get With the Guidelines-Stroke Program. *Circulation.* Oct 12 2010;122(15):1496-1504.

55. Nedeltchev K, Renz N, Karameshev A, et al. Predictors of early mortality after acute ischaemic stroke. *Swiss Med. Wkly.* 2010;140(17-18):254-259.

56. Fonarow GC, Saver JL, Smith EE, et al. Relationship of National Institutes of Health Stroke Scale to 30-Day Mortality in Medicare Beneficiaries With Acute Ischemic Stroke. *J Am Heart Assoc.* Feb 2012;1(1):42-50.

57. Teale EA, Forster A, Munyombwe T, Young JB. A systematic review of case-mix adjustment models for stroke. *Clin. Rehabil.* September 1, 2012 2012;26(9):771-786.

58. Keyhani S, Cheng E, Arling G, et al. Does Inclusion of Stroke Severity in a 30-day Mortality Model Change Standardized Mortality Rates at VA Hospitals. *Circulation. Cardiovascular quality and outcomes.* 2012;5:508-513.

59. Kurth T, Elkind MV. Comparing hospitals on stroke care: The need to account for stroke severity. *JAMA.* 2012;308(3):292-294.

60. Friese CR, Earle CC, Silber JH, Aiken LH. Hospital characteristics, clinical severity, and outcomes for surgical oncology patients. *Surgery.* 5// 2010;147(5):602-609.

61. Rosenberg AL, Hofer TP, Strachan C, Watts CM, Hayward RA. Accepting Critically Ill Transfer Patients: Adverse Effect on a Referral Center's Outcome and Benchmark Measures. *Ann. Intern. Med.* 2003;138(11):882-890.

62. Combes A, Luyt C-E, Trouillet J-L, Chastre J, Gibert C. Adverse effect on a referral intensive care unit's performance of accepting patients transferred from another intensive care unit. *Crit. Care Med.* 2005;33(4):705-710.

63. The Joint Commission. Advanced Certification for Primary Stroke Centers. 2015; http://www.jointcommission.org/certification/primary_stroke_centers.aspx. Accessed January 5, 2015.

64. Kirkham JJ. A comparison of hospital performance with non-ignorable missing covariates: An application to trauma care data. *Stat. Med.* 2008;27(27):5725-5744.

65. Ryan AM, Bao Y. Profiling provider outcome quality for pay-for-performance in the presence of missing data: a simulation approach. *Health Serv. Res.* Apr 2013;48(2 Pt 2):810-825.

66. Reeves MJ, Smith EE, Fonarow GC, et al. Variation and Trends in the Documentation of National Institutes of Health Stroke Scale (NIHSS) in GWTG-Stroke Hospitals. *Submitted to Circulation. Cardiovascular Quality and Outcomes.* 2015.

67. Rubin DB. Inference and missing data. *Biometrika.* December 1, 1976 1976;63(3):581-592.

68. Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychol. Methods.* 2012-09-10 2002;7(2):147-177.

69. Graham JW. Missing Data Analysis: Making It Work in the Real World. *Annu. Rev. Psychol.* 2009;60(1):549-576.

70. Altman DG, Bland JM. Missing data. *BMJ.* 2007;334:424.

71. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med. Care.* August 2013;51(8 Suppl 3):S30-S37.

72. Knol MJ, Janssen KJ, Donders AR, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J. Clin. Epidemiol.* Jul 2010;63(7):728-736.

73. Gorelick MH. Bias arising from missing data in predictive models. *J. Clin. Epidemiol.* Oct 2006;59(10):1115-1123.

74. Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat. Med.* 2003;22(4):545-557.

75. Gomes M, Gutacker N, Bojke C, Street A. Addressing missing data in patient-reported outcome measures (PROMS) - implications for the use of PROMS for comparing provider performance. *Health Econ.* 2015.

76. Hannan EL, Kilburn H, Jr., Lindsey ML, Lewis R. Clinical versus Administrative Data Bases for CABG Surgery: Does it Matter? *Med. Care.* 1992;30(10):892-907.

77. Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SL. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation.* Mar 27 2007;115(12):1518-1527.

78. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare Claims Data to Assess Provider Quality for CABG Surgery: Does It Work Well Enough? *Health Serv. Res.* 1997;31(6):659-678.

79. Shojania KG, Forster AJ. Hospital mortality: when failure is not a good measure of success. *CMAJ.* Jul 15 2008;179(2):153-157.

80. Nicholas LH, Dimick JB, Iwashyna TJ. Do Hospitals Alter Patient Care Effort Allocations under Pay-for-Performance? *Health Serv. Res.* 2011;46(1p1):61-81.

81. Rothberg MB, Pekow PS, Priya A, Lindenauer PK. Variation in Diagnostic Coding of Patients With Pneumonia and Its Association With Hospital Risk-Standardized Mortality Rates: A Cross-sectional Analysis. *Ann. Intern. Med.* 2014;160(6):380-388.

82. Austin PC, Tu JV, Alter DA, Naylor CD. The Impact of Under Coding of Cardiac Severity and Comorbid Diseases on the Accuracy of Hospital Report Cards. *Med. Care.* 2005;43(8):801-809.

83. Goldman LE, Chu PW, Bacchetti P, Kruger J, Bindman A. Effect of Present-on-Admission (POA) Reporting Accuracy on Hospital Performance Assessments Using Risk-Adjusted Mortality. *Health Serv. Res.* Oct 6 2014.

84. Reeves D, Campbell SM, Adams J, Shekelle PG, Kontopantelis E, Roland MO. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Med. Care.* 2007;45:489-496.

85. Reeves MJ, Broderick JP, Frankel M, et al. The Paul Coverdell National Acute Stroke Registry: Initial Results from Four Prototypes. *Am. J. Prev. Med.* 12// 2006;31(6, Supplement 2):S202-S209.

86. Michigan Department of Community Health (MDCH). Michigan Stroke Registry and Quality Improvement Program. 2014; http://www.michigan.gov/mdch/0,1607,7-132-2945_5104_5279_57683-249680--,00.html. Accessed November 26, 2014.

87. Centers for Disease Control and Prevention (CDC). CDC State Heart Disease and Stroke Prevention Programs. 2013; http://www.cdc.gov/DHDSP/programs/stroke_registry.htm. Accessed November 26, 2014.

88. American Hospital Association. AHA Annual Survey Database Fiscal Year 2013. *AHA Data Viewer* 2014; http://www.ahadataviewer.com/book-cd-products/AHA-Survey/. Accessed November 26, 2014.

89. Singer JD. Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics.* 1998;24(4):323-355.

90. Wolf PA, Abbott RD, Kannel WB. Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke.* August 1, 1991 1991;22(8):983-988.

91. Cholesterol, diastolic blood pressure, and stroke: 13 000 strokes in 450 000 people in 45 prospective cohorts. *The Lancet.* 12/30/ 1995;346(8991–8992):1647-1653.

92. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* Oct 2006;59(10):1087-1091.

93. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Stat. Med.* Jul 20 2007;26(16):3057-3077.

94. Janssen KJM, Donders ART, Harrell Jr FE, et al. Missing covariate data in medical research: To impute is better than to ignore. *J. Clin. Epidemiol.* 7// 2010;63(7):721-727.

95. Schafer JL. Multiple imputation: a primer. *Stat. Methods Med. Res.* February 1, 1999 1999;8(1):3-15.

96. Fonarow GC, Smith EE, Reeves MJ, et al. Hospital-Level Variation in Mortality and Rehospitalization for Medicare Beneficiaries With Acute Ischemic Stroke. *Stroke.* January 1, 2011 2011;42(1):159-166.

97. Heckman JJ. Sample Selection Bias as a Specification Error. *Econometrica.* 1979;47(1):153-161.

98. Grotzinger KM, Stuart BC, Ahern F. Assessment and Control of Nonresponse Bias in a Survey of Medicine Use by the Elderly. *Med. Care.* 1994;32(10):989-1003.

99. Clark SJ, Houle B. Validation, Replication, and Sensitivity Testing of Heckman-Type Selection Models to Adjust Estimates of HIV Prevalence. *PLoS ONE.* 10/17/accepted 2014;9(11):e112563.

100. Sales AE, Plomondon ME, Magid DJ, Spertus JA, Rumsfeld JS. Assessing response bias from missing quality of life data: the Heckman method. *Health Qual. Life Outcomes.* 2004;2:49.

101. Stolzenberg RM, Relles AD. Tools for Intuition About Sample Selection Bias and its Correction. *Am. Sociol. Rev.* 1997;62(June):494-507.

102. Winship C, Mare RD. Models for Sample Selection Bias. *Annual Review of Sociology.* 1992;18:327-350.

103. Cuddeback G, Wilson E, Orme JG, Combs-Orme T. Detecting and Statistically Correcting Sample Selection Bias. *Journal of Social Service Research.* 2004;30(3).

104. SAS. The QLIM Procedure. 2015; http://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_qlim_sect001.htm. Accessed March 3, 2015.

105. Drye EE, Normand S-LT, Wang Y, et al. Comparison of Hospital Risk-Standardized Mortality Rates Calculated by Using In-Hospital and 30-Day Models: An Observational Study with Implications for Hospital Profiling. *Ann. Intern. Med.* 2012;256(1):19-26.

106. Sullivan LM, Massaro JM, D'Agostino RB, Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat. Med.* May 30 2004;23(10):1631-1660.

107. Snijders TA, Bosker RJ. *Multilevel Analysis: An Introduction to Basic & Advanced Multilevel Modeling.* 2nd ed. London, UK: Sage Publishers; 2012.

108. Austin PC, Reeves MJ. The Relationship Between the C-Statistic of a Risk-adjustment Model and the Accuracy of Hospital Report Cards: A Monte Carlo Study. *Med. Care.* 2013;00:1-10.

109. Austin PC, Reeves MJ. Effect of Provider Volume on the Accuracy of Hospital Report Cards: A Monte Carlo Study. *Circ Cardiovasc Qual Outcomes.* Mar 11 2014.

110. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Second ed. Thousand Oaks, California: Sage Publications, Inc.; 2002.

111. Sullivan LM, Dukes KA, Losina E. An Introduction to Hierarchical Linear Modelling. *Stat. Med.* 1999;18:855-888.

112. Silber JH, Rosenbaum PR, Brachet TJ, et al. The Hospital Compare Mortality Model and the Volume-Outcomes Relationship. *Health Serv. Res.* 2010;45(5 Pt 1):1148-1167.

113. Clark DE, Hannan EL, Raudenbush SW. Using a hierarchical model to estimate risk-adjusted mortality for hospitals not included in the reference sample. *Health Serv. Res.* Apr 2010;45(2):577-587.

114. Hofer TP, Hayward RA. Identifying Poor-Quality Hospitals : Can Hospital Mortality Rates Detect Quality Problems for Hospitals Identifying Rates Detect Can Hospital Mortality Quality Problems for Medical Diagnoses ? *Med. Care.* 1996;34:737-753.

115. Thomas JW, Hofer TP. Accuracy of Risk-Adjusted Mortality Rate As a Measure of Hospital Quality of Care. *Med. Care.* 1999;37:83-92.

116. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA.* 1999;281(22):2098-2105.

117. Lichtman JH, Leifheit-Limson EC, Jones SB, Wang Y, Goldstein LB. 30-Day risk-standardized mortality and readmission rates after ischemic stroke in critical access hospitals. *Stroke.* Oct 2012;43(10):2741-2747.

118. Saposnik G, Jeerakathil T, Selchen D, et al. Socioeconomic status, hospital volume, and stroke fatality in Canada. *Stroke.* Dec 2008;39(12):3360-3366.

119. Ogbu UC, Slobbe LC, Arah OA, de Bruin A, Stronks K, Westert G. Hospital Stroke Volume and Case-Fatality Revisited. *Med. Care.* 2010;48(2):149-156.

120. Halm EA, Lee C, Chassin MR. Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature. *Ann. Intern. Med.* 2002;137(6):511-520.

121. Birkmeyer JD, Siewers AE, Finlayson EV, et al. Hospital Volume and Surgical Mortality in the United States. *N. Engl. J. Med.* 2002;346(15):1128-1137.

122. Austin PC, Anderson GM. Optimal statistical decisions for hospital report cards. *Med. Decis. Making.* 2005;25:11-19.

123. Keenan PS, Normand S-LT, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circulation. Cardiovascular quality and outcomes.* 2008;1:29-37.

124. Krumholz HM, Lin Z, Drye EE, et al. An Administrative Claims Measure Suitable for Profiling Hospital Performance Based on 30-Day All-Cause Readmission Rates Among Patients with AMI. *Circulation. Cardiovascular Quality and Outcomes.* 2011;4:243-252.

125. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat. Med.* 2006;25(24):4279-4292.

126. Hodgson T, Burke M. On Simulation and the Teaching of Statistics. *Teaching Statistics.* 2000;22(3):91-96.

127. Journal of American Health Information Management Association Staff. Word from Washington: Sights Set on ICD-10-CM/PCS. 2015; http://journal.ahima.org/2015/01/30/word-from-washington-sights-on-icd-10-cmpcs/. Accessed February 2, 2015.

128. Krell RW, Staiger DO, Dimick JB. Reliability of Surgical Outcomes for Predicting Future Hospital Performance. *Med. Care.* 2014;52(6):565-571.

129. Weinstein MC. *Clinical Decision Analysis.* WB Saunders Co.; 1980.

130. DeGroot MH. *Optimal Statistical Decisions.* Wiley; 2004.

131. Austin PC. A comparison of Bayesian methods for profiling hospital performance. *Med. Decis. Making.* 2002;22:163-172.

132. Austin PC. Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Med. Res. Methodol.* 2008;8:30.

133. Bronskill SE, Normand SL, Landrum MB, Rosenheck RA. Longitudinal profiles of health care providers. *Stat. Med.* Apr 30 2002;21(8):1067-1088.

134. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of Databases Designed for Claims Payment versus Clinical Information Systems: Implications for Outcomes Research. *Ann. Intern. Med.* 1993;119(8):844-850.

135. Hammill BG, Curtis LH, Fonarow GC, et al. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circ Cardiovasc Qual Outcomes.* Jan 1 2011;4(1):60-67.

136. Groene O, Kristensen SR, Arah Oa, et al. Feasibility of using administrative data to compare hospital performance in the EU. *International journal for quality in health care : journal of the International Society for Quality in Health Care / ISQua.* 2014;26(S1):108-115.

137. Ding VY, Hubbard Ra, Rutter CM, Simon GE. Assessing the accuracy of profiling methods for identifying top providers: performance of mental health care providers. *Health Services and Outcomes Research Methodology.* 2012;13:1-17.

138. Paddock SM, Adams JL, Hoces de la Guardia F. Better-than-average and worse-than-average hospitals may not significantly differ from average hospitals: an analysis of Medicare Hospital Compare ratings. *BMJ Qual Saf.* Feb 2015;24(2):128-134.

139. Burwell SM. Setting Value Based Payment Goals - HHS Efforts to Improve US Health Care. *N. Engl. J. Med.* 2015;372(10):897-899.

140. Lee TH. Performance Metrics as Drivers of Quality: Getting to Second Gear. *Circulation.* Dec 4 2015;131:967-968.