Essays on the Health Economics of Hospital Quality

by

Shane Michael Murphy

THESIS

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

in the Department of Economics at

Lancaster University

Lancaster, UK

Submitted June 2016

Revisions submitted November 2016

Abstract

This thesis consists of three essays on hospital quality of inpatient care for patients with acute myocardial infarction (AMI) in the United States. First, it explores issues in the measurement of quality, particularly through the estimation of risk-adjusted mortality rates (RAMRs) for hospitals. This work then examines the relationship between hospital quality for AMI patients and the volume of AMI patients.

Chapter 2 proposes using machine-learning techniques, particularly random forests, for risk adjustment of patient severity to predict patient mortality. This work shows that these methods greatly outperform other commonly-used methods in precision of patient risk estimates and also that a facility's estimated RAMR is sensitive to the underlying patient risk-adjustment model.

Chapter 3 asks whether a model which aggregates patient mortality risk for AMI patients matters when estimating RAMRs. To do this, it creates a simulation based on realistic assumptions about how patient case mix can vary by hospital quality and how hospital quality can vary by hospital volume. Because different methods of estimating patient mortality risk have different degrees of precision, the simulation considers variation in this precision and further allows precision to vary by hospital. Again, the ranking of hospitals is sensitive to the method used and this paper finds that common methods are not preferred in many important contexts. Both of the first two chapters pay particular importance to applications of their results to pay-for-performance schemes.

Chapter 4 examines the relationship between quality, measured by RAMR, and volume in hospital health provision for AMI inpatients. The main contribution of the paper is estimate the causal effect of volume on quality. To do this, it uses a novel instrument, the volume of shock and of trauma patients. Previous work has found mixed results and has primarily used the volume of patients with the same condition within a certain radius of the hospital as an instrument for volume within the hospital. This paper argues that this instrument has a number of shortcomings that its instrument does not. This paper tests various specifications used in other work and finds robust results for its conclusion.

Anfangs wollt ich fast verzagen,
Und ich glaubt, ich trüg es nie;
Und ich hab es doch getragen –
Aber fragt mich nur nicht, wie?
– Heine

I put my life in my hands
– Judges

Contents

List of Tables

List of Figures

Acknowledgements

This project has been made possible with the support of many people. My advisors, Professor Colin Green and Professor Bruce Hollingsworth have consistently been there for me since I came to Lancaster University and have supported this work until its finish. Also, two funding bodies have made my PhD at Lancaster University possible: I thank the Northwest Doctoral Training Centre for providing me with a studentship. I also thank the Department of Economics at Lancaster University for travel, training, and data purchasing support.

The staff and faculty of the Department of Economics at Lancaster University have, likewise, been there to support me at every step, particularly Professor Ian Walker, Professor Ivan Paya, Ms. Sarah Ross, and Ms. Caren Wareing. I have also had a number of mentors who taught me many important skills before arriving in Lancaster. I would like to particularly mention Professor George Tolley, Professor Dean Jamison, and Professor Chris Murray.

I would like to thank many of my peers from whom I've learned so much. These include students and faculty at Lancaster University, students at the University of Manchester and the University of Liverpool who are part of the UK's Northwest Doctoral Training Centre, and the many participants of European Health Economics conferences I have interacted with over the past four years. I want to particularly thank the other students in my cohort in the Department of Economics at Lancaster University.

I owe this thesis first and foremost to my partner, both in scholarship and in life, Ayesha Ali. Her patience, thoughtfulness, and creativity have deeply impacted this work. Finally, I would like to thank my family. My parents, my parents-in-law, my sisters, my aunts and uncles, and, again, my wife have been a constant blessing to me and a constant support to my work. In particular, I would like to thank my mother and my mother-in-law, who were with me when I started this process and are with me still.

Declaration of Authorship

I hereby declare that this dissertation and the work presented in it is entirely my own and I have clearly documented all sources and materials used.

This work has not been submitted in any form for the award of a degree at this university or any other institution, nor has this work been published.

Shane Michael Murphy
June 2016

Chapter 1

1.1. Introduction

The most important theme that this thesis considers is the role of empirical measurement in economics. Improving measurement has the obvious value of increased ability to learn about the world. Any foundation, any theory, any hypothesis will have parameters that need to be measured. Among the first questions explored by David Hume's foundational empiricism was the regularity of the sun's rising, but quickly questions faced by a social scientist require increased measurement and statistical sophistication, and measurement became fundamental to empiricism. The applicability of the Poisson distribution quickly became apparent when the number of equestrian accidents was accurately tabulated. Once those tabulations became available, it changed how we think of risk and diversification, for instance.

An important aspect of precision in empirical measurement is the increased efficiency of estimating a statistic when more precise measurements are used. Efficiency is an important goal in econometrics as it plays a role in the reliability of an estimate. Improvements in measurement improve the efficiency of an estimator without sacrificing other properties of an estimator such as bias and consistency. Also, measurement error in a predictor in a regression context, called the errors-in-variables model, leads to biased estimates in the slope of a linear model of the relationship between the predictor and the outcome, known as attenuation bias. This bias can greatly mislead scientists, and should not be ignored.

In applied studies of public policy, empirical measurement plays an important role in the concept of cost-effectiveness, a ratio which can play a role in determining optimal policy. It also is important when measured statistics are used to create incentives, especially in the provision of public goods where competition, price, and markets play a diminished role. An interesting issue arises when considering how measurement of a particular concept is used. It is possible, even likely, that different ways of measuring a concept will be most appropriate in different applications. This can be true at many levels, and an important factor in this consideration is the evaluation of a measurement; how can we determine if one measure is better than another.

It is often possible to determine the value of a measure theoretically. An important case is the estimation of the slope of a linear model of a relationship between a predictor and an outcome in a regression. If an estimate is considered optimal if it is unbiased, then this is a situation where ordinary

least-squares estimation is optimal. However, bias may not be the most important consideration. Before hitting upon the idea of minimizing squared deviations, as is used in ordinary least squares regression, statisticians minimized least-absolute deviation, a procedure which is not unbiased. Even so, least-absolute deviation still has a place in statistics today, as it is more robust to outliers, and is known as median regression. More relevant to this thesis, machine learning techniques are not necessarily unbiased, but they do perform very well in other measures.

In many situations, it becomes more difficult to theoretically differentiate between two different measures or two different statistics. Often a loss function is set up to determine an optimal statistic. In ordinary least squares, the loss function is the square of the residual error of the model. But in measurement, the loss function may be more complex. More interestingly, the loss function may not allow a theoretical determination of which measure to choose. This suggests a second important theme of this thesis; empirical evaluation of the usefulness of a measure or a statistic. Empirical evaluation is especially common when evaluating machine learning techniques. An important aspect of the empirical evaluation of any measure is the uncertainty in its estimate.

Another theme of this thesis, which could have been listed first, is health and health provision. This thesis will pay particular attention to the application of large administrative data sets to measure hospital quality. Administrative data can be aggregated at very large levels; this thesis uses the New York and California State Inpatient Databases from 2005-2007 collected by the Healthcare Costs and Utilization Project, a part of the Agency for Healthcare Research and Quality in the US Department of Health and Human Services. The data set contains approximately 90% of the states' inpatient discharges each year (HCUP 2014) totaling millions of inpatients, although most of the analysis in this paper is based on a subset of patients with acute myocardial infarction, numbering 425,322 inpatients. Data of this size is more possible in health than in other subfields of economics that focus more on private goods.

Patient mortality risk is an issue that plays a central role in each of the chapters of this thesis. Not only are the data large, but there is less theoretical guidance on how to manage the relationships between a patient's age, sex, morbidities, and mortality outcome than there might be in other areas of economics. Both the size of the data and the complexity of the relationships in the data suggest using non-parametric and machine learning techniques (these two terms are not mutually exclusive, machine learning techniques are often themselves non-parametric). Non-parametric techniques are, by their nature, difficult to theoretically evaluate. One important method used to evaluate these techniques and compare between them is to simulate data and compare estimations to simulated

true values. A second method is to randomly split the data into one or more training sets and one or more test sets, and then to use the training set to estimate models and statistics and then compare models based on training sets to results from the test set. In this way, it is possible to empirically evaluate different measures and statistics when computational complexity renders a theoretical evaluation difficult or even impossible.

The final theme of this thesis is understanding quality, in particular quality of patient care by hospitals. This theme brings together issues in empirical measurement and methods in health economics. Economics, especially microeconomics, often focuses on the study of incentives, of scarcity, and of the firm – using monetary profit and output as main variables. Health care is a largely public good provided often by public or non-profit firms, and quality plays a similar role to monetary profit and output for private goods and private firms. The second and third chapters explore empirical measurement and modelling issues around quality and how important those issues are in the incentivizing of high quality hospitals. The final chapter explores the role economies of scale and learning-by-doing play in quality. This thesis was written as three separate chapters. However, the thematic bind between them is very strong. Keeping in mind these themes, the remainder of this introduction will introduce each chapter in more detail.

1.2. Summary

Hospital quality is commonly used for pay-for-performance schemes and in estimating cost-effectiveness of hospital level health interventions. Improving the accuracy of estimation of hospital quality will improve the efficiency of these calculations and is an important goal in health economics and health policy research. Patient outcome, usually patient mortality, is an important measure of quality. Accurate interpretation of average patient mortality at a hospital requires the estimate to be adjusted for the case mix of the hospital. The second chapter considers how machine learning can be used to include more information in the model which will improve risk adjustment of hospital mortality rates. Current methods of risk-adjustment fail to fully incorporate all of the data available in the large administrative data sets used for this purpose. Current research reduces the dimensionality of these data sets by clustering similar comorbidities or selecting comorbidities which best predict mortality. This can cause omitted variable bias and reduces precision in patient mortality risk estimates. This chapter's main contribution is proposing machine learning techniques whereby detailed patient morbidities can be used directly to predict patient mortality risk. Focusing on random

forests, this chapter uses a training set to create a number of different models of patient mortality risk. Compared with models commonly used, it shows that machine learning methods greatly outperform other methods in precision of patient risk estimates. The main application of patient risk is in risk-adjusted mortality rates, which is a statistic that compares a hospital's actual mortality rate to an aggregation of expected patient mortality based on patient risk. The chapter finds that a facility's estimated risk-adjusted mortality rate (RAMR) is sensitive to the underlying risk adjustment model and there is significant variation in the ordering of facilities by RAMRs across models. An important application of this is in the efficiency of pay-for-performance schemes in incentivizing quality. The United States Center for Medicare & Medicaid Services Hospital Value-Based-Purchasing plan is one of the largest pay-for-performance schemes currently in use and includes RAMRs with other quality measures. These rules withhold up to 3% of hospital reimbursements from all included hospitals and reallocate those payments according to the measure of hospital quality (Centers for Medicare & Medicaid Services, HHS 2014).

The third chapter asks whether a model aggregating patient mortality risk among patients with acute myocardial infarction matters when estimating hospital RAMRs. This question is answered by creating a simulation based on realistic assumptions about how patient case mix can vary by hospital quality and how hospital quality can vary by hospital volume. Because different methods of estimating patient mortality risk have different degrees of precision, this simulation considers variation in this precision and further allows precision to vary by hospital. Common methods, including the method used in the 2013 Affordable Care Act's Hospital Value-Based Purchasing Program, start from an estimate of in-hospital mortality or in-hospital predicted mortality where the prediction utilizes a model which includes hospital effects. This measure is standardized by multiplying actual (or predicted) hospital mortality by the ratio of population mortality to predicted mortality of a hospital's patients when the prediction only includes patient risk and does not include hospital-specific effects. This study evaluates different methods of estimating RAMR, shows that the ranking of hospitals is sensitive to the method used and that common methods are not preferred in some contexts, and shows the sensitivity of goodness-of-fit of a RAMR to precision of patient mortality risk estimation. In particular, as precision decreases, standardizing risk adjustment decreases the quality of the estimated RAMR, particularly if the measure is to be used in a pay-for-performance scheme or in choosing the best facility. This paper shows a significant improvement in efficiency of quality estimation, which would improve the efficiency of estimating the determinants of quality, in measuring quality for pay-for-performance schemes, for choosing the best facility to attend as a patient, and for other purposes. This papers most important contribution is the use of a novel simulation framework to evaluate alternative formulae for RAMR.

The fourth chapter combines the patient mortality risk model developed in the second chapter and the hospital RAMR formula evaluated in the third chapter to study the relationship between three important concepts: quality, learning-by-doing, and economies of scale. This chapter looks particularly at provision of care for patients with acute myocardial infarction, which is an important measure used in estimating hospital quality. The main contribution of the paper is the introduction of a new instrument, the volume of shock and of trauma patients. Using this instrument, this paper argues that the between estimates of the relationship between volume and quality represent economies of scale and learning-by-doing. The findings show a much more precisely estimated between effect, when no fixed effects are included, but do give evidence for a within effect as well. This confirms much of the literature, but contradicts recent work by Kim et al. (2016) which did not find an effect in fixed effects models.

These different chapters deal with themes that will be a part of the future of health economics: new tools of data analysis, issues in quality measurement that can shape policy, and how volume effects quality. All three chapters take advantage of large administrative data sets. The second chapter creates a new measure of patient mortality risk which is used in the other two chapters. The third chapter provides a unique evaluation of different methods for aggregating patient risk to hospital RAMRs, an important aspect of quality. While randomized experiments are a gold standard in understanding causal relationships, healthcare provision is an ongoing concern and many important questions cannot be addressed in a lab. This makes the use of data analysis, including instrumental variables as used in the fourth chapter, and other methods, extremely important, especially when discussing issues about quality, where lab conditions may not give appropriate inferences.

Chapter 2

2.1. Introduction

Since Arrow (1963), economists have recognized the difficulty faced by potential patients when judging the expected quality of medical care. In part, this reflects the inherent difficulties in measuring and incentivizing quality of hospital care for those involved in service delivery and for policy makers. Methods to incentivize hospitals to improve quality include pay for performance schemes and providing consumers with report cards on aspects of hospital performance. In the United States, report cards on hospital quality in the United States have been published and used since the mid-1990s. Mortality is a key indicator in any measurement of quality, however comparisons between mortality rates across facilities is confounded by a range of factors, most notably variations in the underlying severity of their case-mixes. The standard approach to these problems is risk adjustment. This measures the composite mortality risk of a hospital's patient case-mix and thereby creates a comparable estimate of hospital or other aggregate level mortality. The main complication faced by current approaches to risk adjustment is dimensionality. This issue results from the large number of potential comorbidities and leads to both model selection and computational problems.

Inpatient data sets may include patient morbidity information that includes all diseases present in the patient during the stay. These are often used to determine reimbursements to the hospital for the patients care. Comorbidity information recorded often uses the International Classification of Disease (Volume 9 – Clinical Modification, used in this paper, is abbreviated ICD-9-CM) categorizations, providing over 15,000 possible comorbidities. Using this information directly creates a very large set of predictors. Sparse, high-dimensional models often do not converge and reducing dimensionality means losing information and leads to the potential for consequential omitted variable bias. The current approaches deal with the problem by grouping comorbidities into a limited number of categories in order to apply generalized least squares estimation of patient mortality likelihood. As a result of the information lost in these categorizations, these methods suffer from high variance and low predictive precision. Another method attempts to select a subset of diagnoses from recorded ICD-9 codes for comorbidities (Pine et al. 2007). This approach will result in omitted variable bias and loss in precision due to the reduced information from reducing the set of comorbidities. This loss in precision results in an inefficient estimate of hospital quality, in effect, overestimating or underestimating quality of a significant number of facilities. Reducing modelling error in mortality prediction is key in establishing the usefulness of risk adjustment for both practical and academic purposes (Smith and Street 2012).

This paper uses machine learning to allow for the inclusion of all information in patient records databases that currently exist to adjust patient mortality risk and estimate hospital risk-adjusted mortality rates. In doing so, this paper demonstrates how this approach overcomes any need to reduce dimensionality and avoids theoretical bounds on convergence faced by logit and probit mortality models. It also shows that these measures are better estimates of hospital risk adjusted mortality rates and the implications of these differences on pay for performance schemes. Focusing on patients with acute myocardial infarction (AMI), this study uses the full detail available in administrative records, greatly increasing the number of included comorbidities. This work is the first to show that risk-adjustment using machine learning can be used with the full detail of patient comorbidities. This paper finds that this detail results in significantly more precise estimates of patient risk.

The usefulness of the model to health economics is expressed with a simple application: the validity of pay-for-performance reimbursements when performance is judged using risk adjusted mortality rates. When these estimates of patient risk are used to estimate hospital risk-adjusted mortality rates, the paper also finds a significant change in how hospitals are ranked according to those rates. If grouped by risk-adjusted mortality rate into quartiles, about one third of hospitals in each quartile using competing methods are placed into a different quartile when using the new method, and about 7% of hospitals in the highest quartile and 7% in the lowest quartile switch places.

2.2. Background

Risk-adjustment is based on comparing the predicted patient outcome with the actual patient outcome. There has been some research in health economics and policy focused on the ability of models to predict cardiovascular patient outcomes (for instance, Grieve et al. 2003), but less frequently with risk-adjustment as the primary goal. Two important aspects of the risk-adjustment process are that the variables used in adjustment have clinical coherence and validity and that the analytical approach takes into account the non-linearity and multilevel organization of the data (Krumholz et al. 2006a). Patient risk is commonly based on patient comorbidity data recorded in administrative discharge abstracts in which hospitals code doctor diagnoses of patients they see. These data can be processed based ICD-9-CM or the ICD 10 scheme associating patient diagnoses with a 3, 4, or 5 digit alpha-numeric code. In this system, multiple diagnoses may be present for a given patient: the first diagnosis coded is the primary diagnosis, and the following are secondary diagnoses.

The total number of secondary diagnoses allowed per patient varies by jurisdiction, but in the US, the payments system is generally structured so only the first eight are used, and thus all states report at least eight secondary diagnoses.

A number of different methods have been used to address the issue of clinical coherence of sample. There are currently over 15,000 diagnoses using ICD-9-CM's scheme, a total that has increased over time. Ellis (2000) suggests that it is intractable to classify individuals at this level of detail, and some effort has been put into reducing the number of possible diagnoses for the purpose of risk adjustment. Elixhauser comorbidity measures categorize the ICD-9-CM diagnoses codes into 30 categories. Logistic regression of mortality on Elixhauser comorbidity measures is a leading risk-adjustment strategy to account for patient-level characteristics in retrospective analyses (Elixhauser et al. 1998), which has been found to be superior to its precedent, the Charlson Comorbidity Index (17 Categories) by Southern et al. (2004), and in Taiwan by Chu et al. (2010). Others risk adjustment schemes include the Adjusted Clinical Groups (32 Categories), Diagnostic Cost Groups/Hierarchical Condition Categories (118+ Categories) used by Kromholz et al. (2006), Clinical Risk Groups (534+ Categories), and Clinical Classification Software (260+ Categories). Hierarchical Condition Categories were found to be better predictors than Charlson and Elixhauser (Li et al. 2010) and is the preferred method of the Center for Medicare & Medicaid Services. Pine, et al. (2007), use forward-stepwise logistic regression as a process of variable selection from the entire set of ICD-9-CM codes.

The total number of categories in many schemes has increased as the number of ICD codes increases. Each of these seeks to provide a clinically coherent set of diagnoses, allowing the universe of diagnoses to be reduced from the number present in ICD, and to reduce the chance that patients with similar symptoms who are diagnosed using similar but slightly different codes are risk-adjusted in similar ways.

An alternative to reducing the number of ICD codes is using classification and prediction algorithms which do not assume a data model and estimate parameters. Tree-based machine leaning methods such as random forest described below can perform this task in a computationally efficient way that avoids estimating parameters for a large number of predictors simultaneously without reducing the dimensionality of the problem through categorization of *a priori* variable selection. Tree based models have a high degree of predictive accuracy, but in contrast to econometric models, are not built to minimize bias. Machine learning methods do not overcome omitted variable bias, endogeneity between institutional quality and coding practices, or endogeneity between patient characteristics and coding practices. However, limitations such as these are present in previously used methods and

thus they do not limit the ability to compare performance between methods. Machine learning methods are increasingly being used in health economics when predictive accuracy is a high priority and the problems of using a method that is not unbiased can be mitigated (Hastie et al. 2009, Lee et al. 2010, Kreif et al. 2015). This paper used the random forest method rather than other machine learning methods such as neural networks or the super learner because tree based methods provide more transparency, as the output includes information about the importance of predictors in the prediction (Breiman, 2003).

An important application of risk adjustment pertains to pay-for-performance (P4P). Risk adjusting performance seeks to incentivize the efficient, quality delivery of care measured with health outcomes. Bird et al. (2005) outline some principals of estimating performance indicators in the health sector. An important factor they note is how the variability in outcomes affects the confidence in estimates of performance. This variability plays a key role in this analysis, and the paper finds that the distribution of risk-adjusted mortality is quite sensitive to the method used to estimate patient risk. Existing research has emphasized the importance of uncertainty in empirical measurement when estimating institutional quality (Goldstein and Spiegelhalter, 1996). Street, 2002 showed that hospital efficiency was sensitive to estimation decisions, and questioned the use of hospital specific point estimates in performance evaluation.

The United States are the Center for Medicare & Medicaid Services (CMS) Hospital Value-Based-Purchasing plan is one of the largest pay-for-performance schemes currently in use and includes risk-adjusted mortality rates with other quality measures. These rules withhold up to 3% of hospital reimbursements from all included hospitals and reallocate those payments according to the measure of hospital quality (Centers for Medicare & Medicaid Services, HHS 2011; Joynt and Jha 2013).

Similarly, adjusting for patient risk in estimating the cost-effectiveness of hospital interventions plays a significant role in the resulting estimates and poor risk adjustment can make a difference when it comes to comparing the cost-effectiveness of interventions. McKay et al. (2008) use privately estimated patient risk values from a model which includes demographic variables and comorbidities to provide estimates of the importance of cost-inefficiency on health outcomes. Cost-inefficiency is estimated using stochastic frontier analysis, and is not found to be an important predictor of hospital performance. On the other hand, Karnon et al. (2013) show that risk adjusting has a great effect in how cost-effectiveness of hospitals is estimated and the ranking of hospital performance in their study of cardiac patients at four hospitals.

Underlying both pay for performance schemes and the cost-effectiveness of hospital interventions is the role that quality and perceived quality play in patient choice of their provider of care. Report cards are a popular way to influence patient perception of hospital quality. There has been a flurry of studies estimating the effect of these report cards, finding that admissions volumes are affected by report cards, especially when report card results are low (indicating high mortality) or below expectations (Cutler et al. 2004, Dranove and Sfekas 2008), but that report cards do not have a great effect on future performance (Dranove et al. 2003). In work with similar results, Wang et al. (2011) points out the importance of patient level over hospital level analysis in this area, while Epstein (2010) notes that reputation about performance plays an important role even where report cards are not available – both in the choices patients make and in how patients are referred by their general practitioners. In all cases, comparing quality among hospitals relies heavily on the quality of the method of risk adjusting mortality rates of the hospitals.

2.3. Methods

The first step in estimating risk-adjusted mortality rates is to estimate mortality risk. Then mortality risk can be used to adjust hospital mortality rates. First, mortality risk is estimated for each patient. Then estimated patient scores are used in a logistic regression mixed model with hospital random effects. The resulting estimates for the hospital effects are then normalized to the underlying mortality rate.

It is common in risk adjustment studies to focus on cases of Acute Myocardial Infarction (AMI). In contrast to many conditions, the protocol for treating AMI is homogenous across facilities and comorbidities are commonly used to estimate patient severity (Antman et al. 2004). Therefore, a patient's mortality is a function of both patient-level characteristics and the facility's adherence to the "best practices" protocol, but not a facility-level choice of treatment strategy. Also, risk-adjusted mortality from AMI has been shown to be correlated at least with risk-adjusted mortality for coronary artery bypass graft, and suggest that hospitals providing good-quality medical management of coronary artery disease also provided a good-quality surgical service (Park et al. 2005).

Hospitals generally collect extremely detailed patient comorbidity data which may be used to estimate severity. In this model, an indicator variable for each disease in this data set is created and a value based on the presence of that comorbidity is assigned to each patient. An indicator variables for basic patient demographic data is also created. Faced with such a large number of predictors, machine

learning techniques provide a way to build models that include as many predictors and interrelations between predictors as necessary (Caruana et al. 2008, Hastie et al. 2009). This model uses a random forest to estimate a patient's mortality likelihood[1] with actual in-hospital mortality as the outcome variable and these comorbidity and demographic indicator as predictors. The fitted outcome of this procedure is *patientscore* which is measured as a percentage of trees for each patient which predict that the patient will die. *Patientscore* is continuous on the interval [0, 1] but should not be thought of as a percentage chance that a patient will die – random forest does not lend itself to this interpretation in the way that probit regression might. The model is generated using half the observations as a training set and the other half as a test set. Model fit statistics are generated for the test set sample, alone, but risk-adjusted mortality rates for facilities are calculated using patient scores from both the test and training set so that there are more observations per hospital. Separating data into a test and training set is common in estimating goodness of fit for machine learning methods to reduce the possibility that goodness of fit scores represent over-fitted models which can result from using the same data to generate and test the model.[2]

Random forests are based on the creation of many classification trees (the general method, classification and regression trees, may be better recognized by its acronym, CART). A classification tree is a binary decision tree that takes a selection of the data at each node and split it into two sets based on its classification according to a variable selected from the set of predictors. Given a chosen predictor, the node splits the observations based on whether or not the observation takes a value for that predictor from a set of values (for instance, if the predictor is binary, then the node splits based on whether that predictor takes a zero or one, if the variable is continuous, the node may split the observations based on whether or not they take a value greater than some constant). At the first node is the entire set of observations being used for a particular tree. At each node, the variable used to split the data is selected based on one of many different splitting criteria. These criteria seek to measure how well the observations are split based on how homogeneous the post-split sets are in terms of the outcome variable of interest. In random forests, the criterion that is usually used (and the criteria used in this paper) is to maximize reduction of Gini impurity (Breiman, 2003). Gini impurity measures of how often a randomly chosen observation from the set of observations at a node would be incorrectly labeled if a label were randomly assigned. This reduction in gini impurity is calculated by looking at a weighted average of the gini impurity of the two subsets that are formed after the

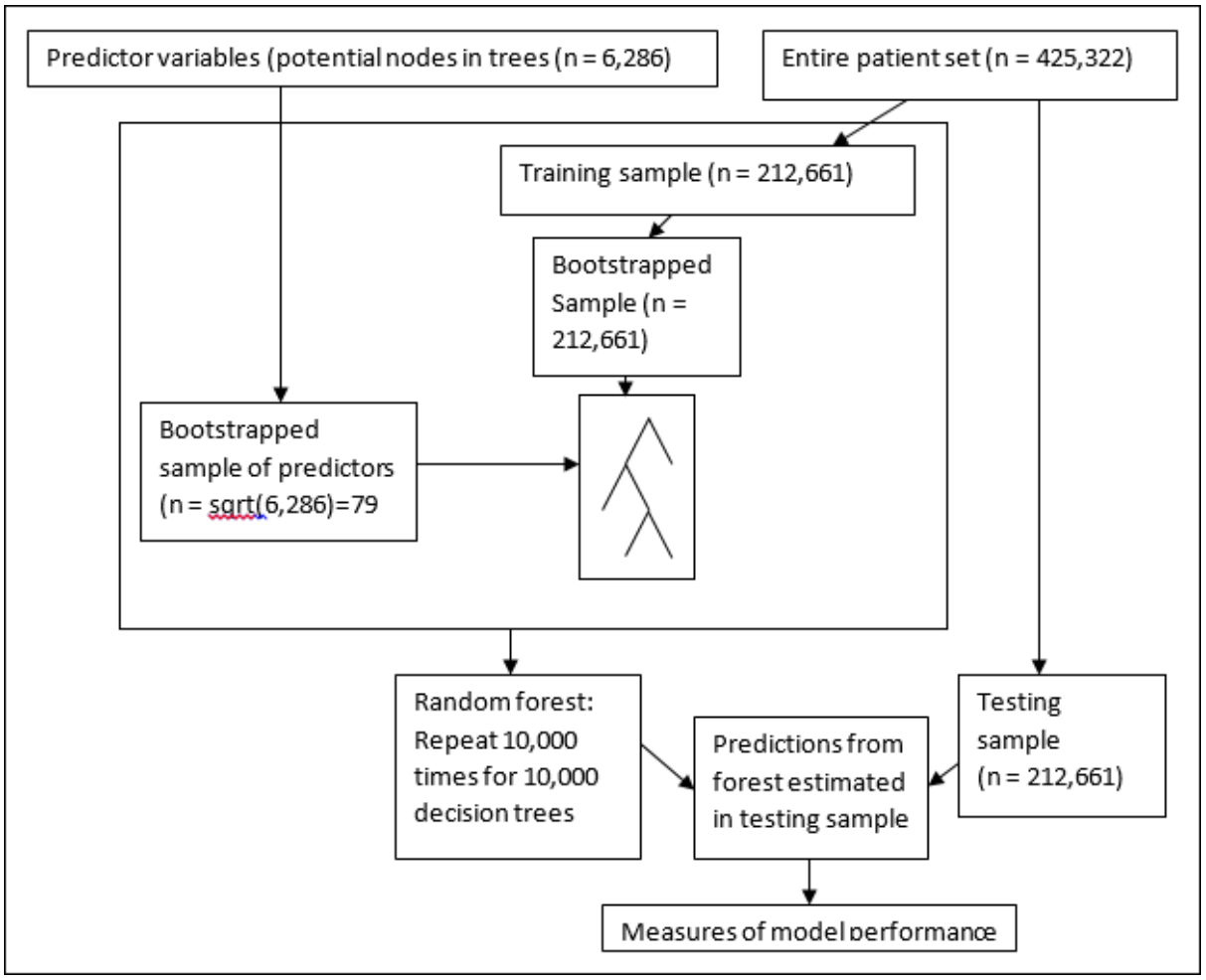
[1] Random forest is computed using R version 3.0.1 and the package RandomForest version 4.7 using default parameters. Alternate runs increasing the number of trees to grow and increasing the number of variables sampled at each node did not change the results.

[2] Great variation in predicted outcomes was not found between a random forest model created using the entire sample compared to one with only a 50% training sample.

observations are split. In random forests, new nodes are added to trees until no further reduction in gini impurity is possible. While generating a decision tree can take many calculations as the gini impurity is estimated for many predictors across many potential nodes, no calculation is computationally complex in the way that matrix manipulation required for generalized least squares regression can be, and thus is well suited for large data sets with many predictors.

Random forest is an ensemble learning method which means that it generates many classifiers and aggregates their results, illustrated in Figure 2.1. These classifiers are individual decision trees, here 10,000 trees are used (changing this by a factor of 10 or 100 did not materially change the results). Trees are generated using a bootstrap sample of the training data and a random subset of the predictors. Each tree is then used to predict whether or not an observation in the test set died, and patient score is based on the percentage of trees which predict mortality for that patient. In some trees it is possible that the key predictors are not selected, and thus much of the predictive power may come from a subset of the trees. This technique is sometimes called bootstrapped aggregating or bagging. Since a subset of predictors is used, each tree is less powerful for predicting than would be a single tree built using the entire set of predictors. However, a tree built from the entire set of predictors may be over-fit, producing a very good predictor on the training set of data but not on the test set of data. Taking a random subset of predictors for each tree is akin to applying a shrinkage method where those predictors are forced to have no effect on a prediction. Thus random forests take on the low variance and high predictive power of shrinkage methods and the robustness to outliers, insensitivity to monotone transformations of predictors, and computability of tree based methods (Hastie T, Tibshirani R, and Friedman J. 2009).

Figure 2.1: Schematic diagram of random forests



For comparison, random forests using the Elixhauser classification scheme as well as performing a logistic regression based model are created. Logistic regression requires an iterative process to converge upon a set of values reflecting the contribution each predictor to the outcome of interest. In a case such as the one here studied, the large number of predictors causes the process to fail to converge, even when the process is allowed a very large number of iterations. To deal with this, a variable selection method is often recommended; Pine, et al. (2007), use forward-stepwise logistic regression as a process of variable selection. This work is replicated using forward-stepwise logistic regression to select a subset of present on admission predictors of equal number (they select 200 predictors) to that selected in Pine et al. Forward-stepwise regression starts with a logistic model with no predictors. It selects from the set of all predictors that predictor which most improves the fit. To choose which predictor most improves fit, at each stage, the predictor that gives the greatest improvement in AIC is selected. Logistic regression on just those 200 predictors is carried out and other predictors are not included in this model. Thus, forward stepwise logistic regression can be

interpreted as a sort of shrinkage method – it can also be interpreted as a method of variable selection as these interpretations are not mutually exclusive.

When predicting binary outcomes, there are many goodness of fit statistics appropriate for estimating model quality (Baldi et al. 2000), and this paper focuses on four such statistics. Let TP be true positives, TN be true negatives, FP be false positives, and FN be false negatives. The statistics considered are: accuracy $\frac{TP+TN}{TP+TN+FP+FN}$, precision $\frac{TP}{TP+FP}$, Matthews correlation coefficient or Phi coefficient $\frac{TP*TN+FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$, and the area under the receiver operator characteristic (ROC) curve. The ROC curve is a plot of the true positive rate or sensitivity of a predictor against the false positive rate or 1-specificity of a predictor where the scores are assigned to outcomes dependent on their comparison to some threshold (scores above the threshold are assigned mortality equal one, below are assigned zero). The area under the ROC curve is equivalent to the probability that a randomly chosen estimate from the set of observations with mortality equal to one will be scored higher than a randomly chosen estimate from the set of observations with mortality equal to zero. Since *patientscore* is not a prediction of mortality, but the percentage of the trees in the random forest that predict mortality, for these statistics to be calculated a threshold value on must be chosen between zero and one such that the patient is considered to have been predicted to die if their *patientscore* is above the threshold and to have survived if their *patientscore* is below the threshold. For each method, a number of thresholds may be chosen, and a threshold such that the estimated mortality rate will equal the true mortality rate is chosen in this paper, although a threshold such that the MCC is maximized is sometimes recommended as well (Baldi et al. 2000).

Risk-adjusted mortality rates for hospitals are estimated from patient scores by estimating a logistic regression[3] using actual mortality (*patientOutcome)* as a binary outcome variable, the logit transformation of patient scores (*patientScore*) as a fixed effect, and the individual hospital as a random effect:

(1) $patientOutcome = \alpha * logit(patientScore) + \gamma_{hosp} + \epsilon.$

Where γ_{hosp} represents the relationship between the patient outcome and a hospital (so $hosp \in Set\ of\ Hospitals$) and is parameterized as a random variable (Gelman, 2006):

[3] Logistic regression is performed using computed using R version 3.0.1 and the package lme4 version 1.6 using default parameters.

$$\gamma_{hosp} \sim N\left(0, \sigma^2_{hosp}\right).$$

σ_{hosp} is the standard deviation of model errors at the hospital level. Although patient score is not a probability measure, *logit(patientScore)* is used to improve the interpretability of this hospital measure. In particular, risk-adjusted mortality rate (*RAMR*) for a hospital can be estimated by the inverse logit of the conditional modes of the random effects given the observed data values and the estimated parameter values ($\tilde{\gamma}_{hosp}$) plus the logit of the mean mortality rate across the entire population (11.7%):

(2) $RAMR = invlogit\left(logit(.117) + \tilde{\gamma}_{hosp}\right)$

2.4. Data

The New York and California State Inpatient Databases collected by the Healthcare Costs and Utilization Project contains approximately 90% of the state's inpatient discharges each year (HCUP 2014). The data collected is primarily used to calculate reimbursements for patient care. Variables include the patient's age, sex and a series of ICD-9-CM diagnosis codes. States vary in the number of diagnoses recorded per patient per admission, but the minimum value across all states is nine. The first code listed is designated as the primary diagnosis. A secondary diagnosis is defined as one located in positions two through nine. At most the first nine diagnoses are used for each patient to ensure comparability across states. Each ICD-9-CM code consists of three to five digits. The first three digits represent a broad category of diagnosis. The fourth and fifth digits (if they exist) represent the sub-category and sub-classification of the diagnosis, respectively. For this work, all comorbidities which have the same first three digits are grouped, this is compared to results with those where all five digits are used. The first three digits allow diagnoses to be grouped in clinically coherent ways (Krumholz et al. 2006b).

The sample is limited to cases between 2005 and 2007 in which Acute Myocardial Infarction (AMI) is either the primary or a secondary diagnosis. Further, hospitals with fewer than 25 cases are dropped, as recommended by CMS (Grady et al. 2013). The sample consists of 425,322 patient-level observations, an average of 220 patients (sd = 190) per hospital and the mean hospital mortality rate is 11% (sd = 0.51).[4] Useful demographic variables such as the age and sex of the patient, as well as the

[4] Across New York and California, 4.3% of all observations which listed AMI did so at the tenth or greater position. Those observations were excluded from this analysis. Their inclusion did not materially change any of the reported results.

primary and eight secondary diagnoses can be used to create risk adjustment models from hospital administrative data. In the ICD-9-CM comorbidity coding schemes, the fifth digit, where present is a sub-classification code, and fourth digit is a sub-category code. A 3-digit ICD-9 code is created by dropping the fourth and fifth digit, leaving three digit alpha-numeric codes representing broad categories of diagnoses, which while broad, are more detailed than those of other categorization schemes. ICD-10 codes, which have been used in the US since October 2013, are not perfectly resolvable into ICD-9-CM codes, but are similar and may be used in the same way as ICD-9 codes. Age is binned into five year intervals, and along with a dummy for the patient sex and five dummies for different patient admission types along with 936 3 digit ICD-9-CM based binary comorbidity indicators totaling 987 predictors, all of which are binary variables. It is unnecessary to control for the non-linear effect of age on mortality in tree based models, and has no effect on the results, but is done for comparability to other research.

An important issue facing researchers is the choice between risk adjustment as a prospective task versus a retrospective task. Prospective models rely more on information related to chronic conditions that persist over time, while retrospective models attach relatively more weight to information related to acute conditions (Ellis 2000). As a retrospective task, risk adjustment uses all information revealed about the patient through the care process. This involves including all comorbidity diagnoses given to a patient, including potentially conditions which the patient receives due to poor quality care, in-hospital infections, or accidents that occur during the treatment process. These issues make prospective risk adjustment more appealing. Another argument for preferring prospective risk adjustment is that patient selection can be performed only from prospective information. Patient selection is an important consideration if there is concern that moral hazard affects hospital patient mix and thus could affect this analysis. Dunn et al. (1996) finds retrospective models greatly outperform prospective models in estimating patient mortality risk. However, in estimating facility performance it may be that prospective models are preferred. Data limitations minimize the ability to perform prospective analysis, however the data includes a variable indicating if a comorbidity is present upon admission to the hospital. When only those comorbidities that are present upon admission are used, estimated patient risk is more prospective than when all comorbidities are used. The possibility that the present on admission variable is not used uniformly across hospitals and the possibility of gaming this variable limit the applicability of these estimates. Using all comorbidities is briefly compared to using only those present on admission and in the preferred model only comorbidities present on admission are used.

2.5. Results

Patient score summary results for different sets of predictors are reported in Table 2.1. Reported results are generated from out-of-sample observations only, and they are robust to alternative parameters of the random forest model.[5] The principal statistic of interest is the area under the ROC curve. It is equivalent to the c-statistic, which represents the probability that a randomly-chosen patient who died will have a larger score than a randomly-chosen patient who survived. To estimate accuracy and precision of a prediction, the analysis must include a choice of cutoff, a score above which the patient will be predicted as dead and below which the patient will be predicted as alive. Since the proportion of patients expected to survive is known, a choice of that cutoff is made so that the percentage of patients predicted to survive is equal to the actual percentage. Another measure of quality of prediction is the Matthews correlation coefficient, which combines accuracy and precision into one measure.[6]

[5] Neural networks was also attempted and found similar results. This paper focusses on random forests.

[6] A cutoff such that patients with scores below the threshold are predicted to live, and those with scores above are predicted to die is used (this paper's cutoff equaled the mortality rate observed in the data, or 11.2%). The cutoff can also be varied to calculate the maximum value possible for Matthews correlation coefficient, accuracy and precision. All these values are reported in Table 1.

Table 2.1: Random forest model results for each diagnostic information set

Model	*i*	*ii*	*iii*	*iv*	*v*
Diagnostic	Elixhauser	3-Digit POA	Stepwise	All 3-Digit	All 5-Digit
ICD-9-CM	-	3-digit	5-digit	3-digit	5-digit
Number of predictors	81	987	200	1,011	6,286
Training set	50%	50%	50%	50%	50%
C-Statistic	0.712	0.899	0.853	0.876	0.909
Matthews correlation coefficient	0.146	0.264	0.230	0.234	0.271
Accuracy	0.901	0.912	0.910	0.910	0.912
Precision	0.397	0.842	0.696	0.720	0.860

Note: Patient scores are generated using a random forest with different sets of predictors for each model. Results are generated from out-of-sample observations only. MCC, accuracy, and precision are calculated with the fraction of deaths set at the rate observed in this subset of the data (11.2%). Due to the high number of cases, bootstrapped standard errors estimated on each statistic were very low (below 0.005). Data is from California and New York State Inpatient Database, 2005-2007.

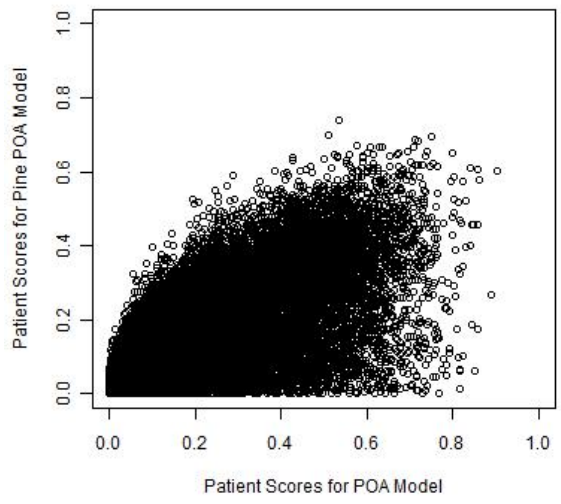
Focusing on c-statistic, models using present on admission (POA) 3-digit ICD-9 codes (*ii*) and using all 3-digit ICD-9 codes (*iv*) perform similarly. To test the importance of the final two digits, a model using all 5-digit ICD-9 codes is estimated (*v*). The 5-digit ICD-9 code model performed similarly to the 3-digit codes, but this result may depend on the size of the training set. All three significantly outperform the models proposed by Elixhauser *et al.* (1998) (*i*) and an approximation of the model used in Pine *et al.* (2007) (*iii*). Very little distinction between models using 3-digit and 5-digit ICD-9 codes is found, even though the latter uses the full set of 15,000 5-digit ICD-9 codes. This increase in predictors is computationally more intensive (which is why its training set is smaller). This implies the informational value of the fourth and five digits is relatively small. All of these statistics are generated using the test set only, but the training set gives very similar estimates, as might be expected given the large number of observations.

It is assumed that including only diagnoses that are POA represents the ideal measure of exogenous patient-level risk. Figure 2.2 illustrates scatters of patient-level scores, and it shows a strong correspondence between models using only POA diagnoses (*ii*) and all diagnoses (*iv*). Table 2.2 is a correlation matrix of facility-level scores generated from each of the five risk-adjustment models with own model standard deviations down the triangle.[7] It reports both the Pearson correlation of facility-level quality scores and the Spearman correlation of the rankings that result from those scores. Comparing models with POA diagnoses (*ii*) to all diagnoses (*iv*), the facility correlations are extremely high: 0.948 for the scores and 0.953 for the ranking. This suggests that in jurisdictions without POA coding, the closest proxy is to use all available diagnosis codes. The diagonal of this table are the standard deviation of hospital rankings within that model. This statistic expresses how the much the variation in patient scores affects the distribution of hospital risk-adjusted mortality rates. Since this method models those rates using a Gaussian-distributed random effect with expected mean equal to the underlying rate of mortality in the data, the standard deviation of these scores is the key statistic in understanding variation in this distribution. An f-test shows that pairwise difference between these standard deviations is significant in all cases where the method using random forest on 3-digit present on admission predictors is used, but not in any of the other cases. This significance suggests that the presence and accuracy of present on admission indicators in patient data may be an important consideration when choosing an optimal risk adjustment process.

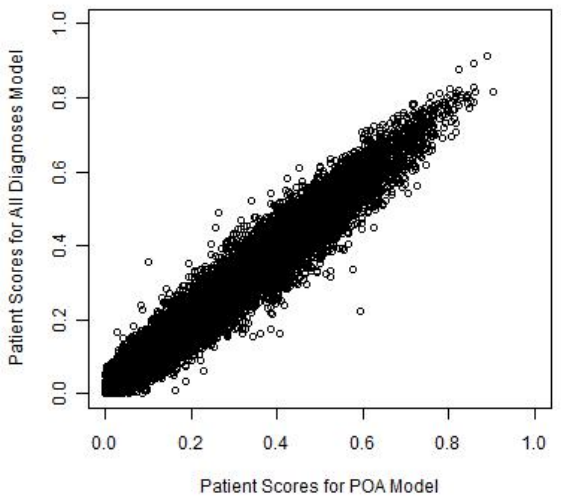
[7] To increase the number of patients per hospital, both in-sample and out-of-sample observations are used in these regressions.

Figure 2.2: Scatterplot of Patient Scores across Models.

Panel A: Model POA 3-Digit ICD-9 Diagnoses (*ii*) versus POA Pine 2007 Inspired mode (*iii)*



Panel B: Model with POA 3-Digit ICD-9 Diagnoses (*ii*) versus All 3-Digit ICD-9 Diagnoses (*iv*)



Note: Patient scores are generated using a random forest with different sets of predictors for each model. Data is from California and New York State Inpatient Database, 2005-2007.

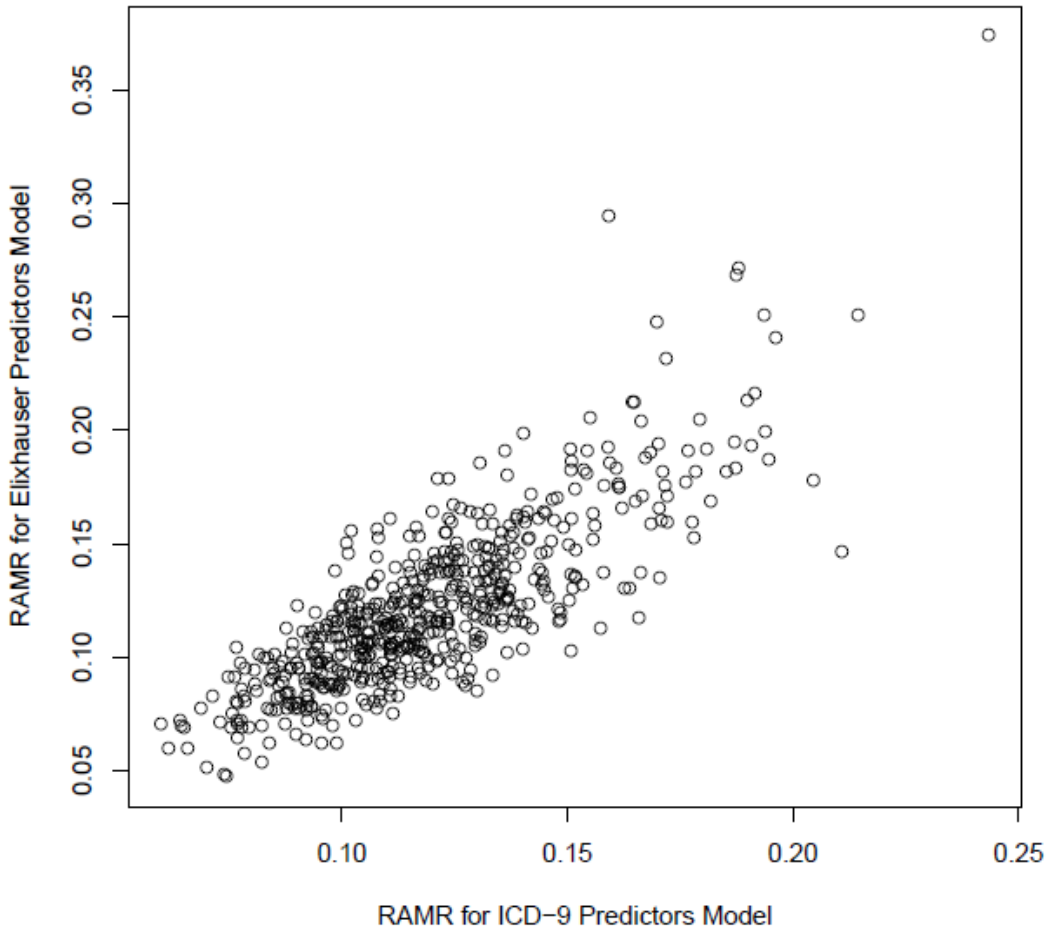
Table 2.2: Pearson and Spearman Correlation Matrix of facility-level quality scores generated by each of the five diagnostic models with own model standard deviations down the diagonal

Statistic	Model		*I* Elixhauser	*ii* 3-Digit POA	*iii* Stepwise	*iv* All 3-Digit	*v* All 5-Digit
Pearson correlation of facility-level quality scores	*i*	Elixhauser	0.237	0.755	0.859	0.827	0.737
	Ii	3-Digit POA		0.296	0.906	0.948	0.910
	Iii	Stepwise			0.234	0.967	0.855
	Iv	All 3-Digit				0.238	0.889
	V	All 5-Digit					0.223
Statistic			Elixhauser	3-Digit POA	Stepwise	All 3-Digit	All 5-Digit
Spearman's ρ correlation of the ranking of facilities by quality score.	*i*	Elixhauser	0.237	0.779	0.852	0.832	0.740
	ii	3-Digit POA		0.296	0.917	0.953	0.898
	Iii	Stepwise			0.234	0.959	0.849
	iv	All 3-Digit				0.238	0.882
	v	All 5-Digit					0.223

Note: Correlation matrix presents correlation between hospital scores across methods, with diagonal representing the standard deviation of hospital scores. Hospital Risk-adjusted Mortality Rates (RAMR) estimated from a hospital random effect on a logistic model with AMI patient mortality as the outcome variable and patient score as the input variable. Patient scores generated using stepwise logistic regression on 5-digit predictors generating a logistic model with 200 predictors for the stepwise model. Patient scores are generated using a random forest with different sets of predictors for all other models. Data is from California and New York State Inpatient Database, 2005-2007.

To further evaluate the usefulness of this metric, hospital risk-adjusted mortality rates is compared between using 3 digit ICD-9 predictors to the hospital risk-adjusted mortality rates using Elixhauser predictors. Figure 2.3 is a scatter plot of hospital scores for the 583 hospitals in this New York and California dataset across these two methods. Since 2013, CMS has withheld a percentage of payments to US hospitals to fund the Hospital Value-Based Purchasing Program. In 2013, the first year of this program, this was 1% of payments. This is scaling up, rising to 2% by 2017 and continuing at that level thereafter. These funds are awarded to hospitals based on a rank ordering of quality of care as measured through process indicators, outcome indicators, and patient experience indicators. Payments for performance in 2013 were announced in November of that year; approximately half of hospitals would see payments change between -0.2% and +0.2%, with a quarter being above that level and a quarter below. The half with minimal change is defined by the CMS as breaking even through the program (Conway 2013). If quality were entirely based on risk-adjusted mortality rates using the 3-digit POA method (model ii), about one third of hospitals losing over 0.2% of payments using the Elixhauser method would now be defined as breaking even; and about 7% of hospitals losing over 0.2% would instead gain over 0.2%. This magnitude of change is present for hospitals estimated to be of high quality using the Elixhauser method, about a third of hospitals gaining over 0.2% would merely break even using this method and about 8% of hospitals gaining over 0.2% are would instead lose over 0.2%. Thus the method strongly affects a significant amount of hospital payments. Rosenthaul and Frank (2006) review existing pay for performance programs in health and find that they have limited effectiveness. They suggest that the small size of the bonus payments plays a role, to which this analysis adds the inefficiency of currently used estimates of facility performance.

Figure 2.3: Comparing Hospital Risk-adjusted Mortality Rates



Note: Hospital Risk-adjusted Mortality Rates (RAMR) estimated from a hospital random effect on a logistic model with AMI patient mortality as the outcome variable and patient score as the input variable. Patient scores are generated using a random forest with different sets of predictors for each model. Data is from California and New York State Inpatient Database, 2005-2007.

An important existing procedure for risk adjustment using detailed comorbidity information is forward-stepwise logistic regression as shown in Pine, et al. (2007). A similar c-statistic and accuracy is found, but significantly improved precision. This suggests that in this case forward-stepwise logistic regression does not overcome issues related to bias compared to tree-based methods.

In order to express uncertainty in hospital specific quality estimates, O used hospital-clustered bootstrapping to generate 500 datasets from the test-set admissions. I then estimated hospital risk-adjusted mortality rate for each bootstrapped dataset and ranked facilities for that bootstrapped sample. I then calculated across all bootstraps each facilities mean absolute change in ranking between that method and using the Elixhauser method for each hospital and 95% confidence interval around that ranking. These means were then normalized so that for each hospital, the mean difference is set to be positive. Thus, a negative value for the 2.5 percentile implies that at least 2.5 percent of the time the hospital ranking moved in the opposite direction of the mean. Using 5-digit ICD-9 codes gives a mean absolute change of 5.8 (a mean 2.5% of 0.1; and a mean 97.5% of 11.0), while 3-digit POA ICD-9 codes gives a mean change of 2.9 (-0.2; 5.5) and all 3-digit ICD-9 codes gives a mean change of 7.1 (1.1; 14.5). Thus, we can be reasonably certain that the variation between rankings comparing the new methods with Elixhauser is more than variation that would come from random chance.

2.6. Discussion

This paper assumes that morbidities in this data are coded to represent the best belief of the diagnosing physician. However, this may not be true and instead, coding may be performed to optimize patient risk to give highest estimate of hospital quality. This is unlikely, not least because the data is collected for the purpose of reimbursements and not for quality estimation and because patient care depends in part on accurate recording of diagnoses. However, this gaming could occur through recording patient comorbidities that inflate a patient's mortality risk. Any model used will necessarily associate certain diagnoses with a higher predicted mortality than other similar conditions. In recording comorbidities, doctors and health administrators sometimes must choose between very similar conditions. In such cases, allowing great specificity in conditions may allow doctors or health administrators to choose a condition which a model gives higher weight in predicting mortality. Using a non-parametric model such as random forests lessens the ease of taking this action, but does not completely rule it out. Regardless of whether such a patient survives, this type of gaming will result in a better risk-adjusted mortality rate for a hospital than they would otherwise receive. Unfortunately, it may not be possible to prevent this type of gaming in all cases. However, if it is known that switching

is occurring between two similar conditions, those conditions can be combined into one condition for the purposes of the model. Also, using aggregated ICD codes (in the case of ICD-9 CM, this is done by using only the first 3 digits) minimizes the chances that a condition could be included in the model under two different codings.

Another issue that is true for all in-hospital mortality based measures of quality is that patients who are likely to die soon could be discharged to avoid their inclusion in in-hospital mortality data. For this reason, 30-day mortality, which includes mortality within 30 days of discharge from the hospital, is commonly used instead of in-hospital mortality. 30-day mortality is not included in this data set and not used in this paper.

A third issue is related to the cross-sectional nature of the data. Random forests as used in this paper are not suited to, for instance, survival analysis. However, recent work has extended machine learning methods and even random forests to such data and has found similar improvements to prediction (Ishwaran et al, 2008). Further research of the problems in this paper should link patients visits so that information from prior visits can be included in prediction, and may extend on that work.

2.7. Conclusion

Our results show significant improvement in risk-estimation for the purposes of risk-adjustment when using random forests on all ICD-9 codes over current methods. The paper also shows that facility rankings vary significantly between the preferred risk-adjustment model and models which use more limited sets of patient morbidity information such as variable selection models such as forward stepwise logistic regression and variable categorization methods such as using Elixhauser's Comorbidity Index. Variation is extremely small between a risk adjustment model using POA diagnosis codes and a model which uses all available codes. Therefore, if one's objective is to generate facility-level quality scores in jurisdictions in which POA codes are unavailable, then the preferred risk-adjustment strategy is to use all available information. This conclusion is conditional on a risk-adjustment strategy that utilizes machine learning and large volume of data, and one may draw different conclusions under different conditions. However, given rapid increases in both the availability of data and computing power, these conditions may be widely applicable to contemporary health economics research.

This paper has shown that risk adjustment using with all ICD-9 codes directly in the analysis can be performed using random forests. Future research to confirm these results should also consider using disease classification from volume 10 of the International Classification of Diseases as well as

considering its application to risk adjusted readmission rates. It has also shown that such a method greatly improves the predictability of the model over GLS methods currently common in the literature and used in government and industry applications. It has also shown that this improvement in predictability results in significantly different estimates of hospital quality and that pay-for-performance schemes and quality reports would reflect this difference. In particular, this paper shows that the quality of as many as one-third of hospitals would be mis-categorized in a scheme which uses RAMR to group facilities into three groups: below average, average, and above average. As such, using a machine learning method such as random forests using all POA ICD-9 codes would be the preferable method for risk adjusting patient mortality.

Chapter 3

3.1. Introduction

The Affordable Care Act's Hospital Value-Based Purchasing Program is a program which began in 2013 and withholds a portion of a hospitals federal reimbursements and distributes those funds to hospitals in proportion to the hospital's quality ranking. The ranking is based on a number of measures, including risk-adjusted mortality rates (RAMRs) and risk-adjusted readmission rates. This type of program is often called a pay-for-performance program and represents a large financial incentive to hospitals to perform well in these measures. Research on measurement issues in health economics has considered the bias that might come in estimation due to reporting heterogeneity (Bago d'Uva et al. 2008). Less focus has been given to the question of how important the method of aggregation of a statistic in health economics is to the usefulness of the resulting estimate. This paper addresses the question of how to calculate RAMRs. This paper considers different health economics applications and suggests the preferred method of estimating RAMRs.

In-hospital patient mortality risk can be estimated upon patient arrival with increasing accuracy and precision due to increased detail and regularity in administrative coding of patient morbidities upon arrival and the use of non-parametric methods to capitalize upon non-linearities and interaction between different morbidities. A hospital risk-adjusted mortality rate (RAMR) is an estimate of the risk of mortality at a hospital that controls for uneven case mixes at different hospitals and is used as an estimate of hospital quality. There has been some research on the role of the estimation of patient mortality risk in estimating RAMR (Pine 2007, CMS; HHS 2011). Less focus has been paid to the question of how to use patient risk to adjust mortality rates. While the method used to estimate patient risk and the method used to apply those risk estimates to adjust mortality rates was debated during the creation of the Affordable Care Act passed in the United States in 2010 (CMS; HHS 2014), there does not exist metrics to compare the usefulness of methods used to estimate RAMRs from patient risk. This paper proposes such metrics and argues that the choice of formula used to estimate RAMR has a significant effect on hospital scoring and relative ranking.

This paper identifies a number of competing formula for RAMR as a function of patient risk and realized patient mortality. It focuses on three important uses of this measure: (1) by patients who may choose between different facilities on the basis of quality, (2) by policy makers who seek to reward high quality facilities through pay-for-performance schemes and to close low quality facilities, and (3) by economists and health policy researchers who wish to use RAMRs in their analyses, either by estimating the determinants of RAMRs (that is RAMR as an outcome variable) or by using RAMRs as a

predictor in an analysis of another outcome. For each of these applications, three tests can be used to compare the accuracy of a formula in estimating RAMR. Simulation from the distribution of patient risk and RAMRs among Acute Myocardial Infarction (AMI) patients in New York hospitals from 2005-2007 is used to estimate the quality of each formula and provide preferred formulae each application.

This paper makes three contributions. First, in order to assess the relative performance of the different formula, this paper proposes a taxonomy of RAMR formula. Second, it designs a simulation framework to model patient risk and hospital quality. Third, it proposes different tests to estimate the usefulness of a risk-adjustment formula to achieve various health economics tasks. Given a set of patient mortality risks, the most intuitive and possibly the most common formula for RAMR is to multiply a hospital's excess mortality ratio by the underlying mortality rate in the population (Guru et al. 2008; Hannan et al. 2013). Another method involves replacing the actual hospital mortality with the expected value of a hospital's mortality from a "random effects" model of hospital mortality in the estimation of a hospital's excess mortality ratio. This method is recommended for use in the 2013 Affordable Care Act, and is not uncommon elsewhere in the literature (Krumholz et al. 2006b; Grudy et al. 2013). Since the hospital specific effect is estimated using a probability function which can shrink extreme values towards the mean (commonly called a random effect in economics), this process can smooth RAMR estimates when a hospital has fewer patients. Alternatively, a hospital specific effect is sometimes used directly as a control variable (see Mark et al. 2005).

3.2. Background

In one type of formula to estimate RAMR, first a binary outcome model of patient mortality as a function of patient mortality risk is created. This model gives an estimate of the probability of patient mortality as a function of patient mortality risk independent of the hospital the patient attends. The sum of these estimates of patient mortality risks for all patients at a given hospital is an estimate of the hospital's risk-adjusted expected mortality rate. This model is compared to an estimate of the actual mortality rate in a hospital. There is a common alternative to obvious method to estimate actual mortality of dividing total deaths by total patients. The alternative is to use estimate a multilevel binary outcome model which includes hospital-specific effects and use the hospital-specific effects as the estimate of the actual mortality rate. The ratio between the estimated mortality rate independent of the hospital and the estimated actual mortality rate is sometimes called an excess hospital mortality ratio. Multiplying this ratio by a population average mortality rate gives a standardized risk-adjusted mortality rate.

There has been limited research on the formula used to estimate RAMR. Silber et al. (2010) study the model recommended by Medicare's "Hospital Compare", which is equivalent to that used by the Affordable Care Act, to estimate annual RAMR for AMI cases. They show that the accuracy of the model suffers because some hospitals do not receive very many patients per year. The paper notes that hospital volume is known to effect hospital quality and suggests that estimated RAMRs would be more accurate if other information, particularly volume, is used in the estimation. This paper also notes the advantage of using random effects in modelling RAMRs using a multilevel model because it provides shrinkage of the estimated rates towards the mean across all hospitals (Gelman, 2006), an effect which results fewer estimates of extremely high and extremely low performing hospitals, particularly for hospitals with fewer patients.

Simulation is frequently used in health research and has been used to address issues with hospital quality measurement. Hofer & Hayward (1996) provide an early example of the use of Monte Carlo simulation to understand if simulated variation in hospital quality due to different rates of preventable mortality would be detected in variation in mortality rate. Even in cases where they assume perfect case-mix adjustment, where only patients' primary diagnoses are considered and no risk adjustment is performed based on additional information about patient case-mix, they find that detection of variation in mortality rates may not be possible. More recently, Koetsier et al. (2012) use simulation of RAMRs to see how different sequential stopping tests used in quality control perform at the task of finding an unexpected decrease in hospital quality over time. Their paper considered a case where hospital quality is measured repeatedly over time and after a given period, the mortality rate increases slightly (representing a decrease in hospital quality). It then asks how many additional periods of measurement at this lower quality are required before different sequential stopping tests signal a change in the RAMR.

To understand how one might measure the usefulness of an RAMR formula, it is important to note some of the applications of hospital RAMRs. Important applications of risk-adjustment include pay-for-performance (P4P), the role of perceived hospital quality in patient choice, and estimation of the relationship between quality and its determinants. Risk-adjusting performance seeks to incentivize quality delivery of care independent of cost and using health outcomes as the dependent outcome (Ellis 2000). Bird et al. (2005) outline some principals of estimating performance indicators in the health sector. An important factor they note is variations in patient case-mixes across facilities. This variability plays a key role in this analysis and an explicit attempt to model systematic and random case-mix heterogeneity across facilities. The two most prominent P4P plans introduced to date in the United States are the Center for Medicare & Medicaid Services (CMS) Hospital Value-Based-

Purchasing plan and Hospital Readmissions Reduction Program which use risk-adjusted 30-day readmission and risk-adjusted 30-day mortality as measures of quality. Adjustment is performed by categorizing comorbidities using a modified set of CMS HCC indicators. Collectively these programs reallocate up to five percent of DRG-based reimbursements for inpatient stays (Centers for Medicare & Medicaid Services, HHS 2011; Joynt and Jha 2013). The value of pay-for-performance schemes relies crucially on the rank correlation (i.e. Spearman correlation) between estimated RAMR and the hospital's actual RAMR. As with many of the analyses discussed in this review, this work uses the in-hospital mortality instead of the very highly correlated 30-day mortality.

Another important consideration in the importance of accurate RAMR estimation is the role that quality and perceived quality play in patient choice of their provider of care. Report cards are a popular way to influence patient perception of hospital quality. There has been a flurry of studies estimating the effect of these report cards, finding that admissions volumes are effected by report cards, especially when report card results are low (indicating high mortality) or below expectations (Cutler et al. 2004, Dranove and Sfekas 2008), but that report cards do not have a significant effect on future performance (Dranove et al. 2003).

Similarly, hospital mortality rate and RAMR is used to estimate the importance of hospital level variables. In the health economics literature, Mark et al. (2005) use risk-adjusted mortality rates as an outcome variable in a study of HMO penetration in the US. More recently, risk-adjusted mortality rates were treated as a control in McKay et al. (2008), who estimate a model of observed mortality on cost inefficiency, patient case mix, hospital quality, and hospital volume. Volume is an extremely common predictor of interest, Finks (2011) look at how the relationship between volume and RAMR has changed over time. Inaccurate estimation of RAMR can be interpreted as measurement error plays an important role in the biasedness of coefficient and standard error estimates in regressions involving RAMR such as these. This study explicitly simulates the relationship between volume and RAMR, but does not seek to provide a causal estimate of this relationship.

RMSE, bias, correlation, bias2

This paper assumes a true RAMR and considers the usefulness of different models of RAMR comparing modeled and true RAMR using four metrics, Root mean squared error (RMSE) between true and modeled RAMRs, bias between true and modeled RAMRs, correlation between the true and modeled RAMRs, and bias in the estimation of a relationship between RAMR and a correlated predictor between true and modeled RAMRs. RMSE is intuitively similar to sum of squared errors (SSE). Minimizing RMSE is equivalent to minimizing SSE, the optimization condition of ordinary least squares

regression. Unlike SSE, RMSE has the same scale as the data, and thus gives us a good idea how wrong, on average, an estimated RAMR is. The closer RMSE is to zero, the better. Bias estimates the direction, on average, of the error in estimation of RAMR. A high correlation of estimated RAMR to actual RAMR is extremely important for the credibility of a pay-for-performance program that emphasizes relative position of ranking of hospitals. Bias in estimation of $\beta_{(3)}$ provides a flavor of how important accurate estimation of RAMR is for scholars seeking to estimate the determinants of hospital quality and is described in more detail later.

3.3. Data

Inpatient data on patients with primary diagnosis of acute myocardial infarction (AMI) is frequently used for work on RAMR. A primary reason is that mortality is the outcome of overwhelming importance to the treating doctor, rather than, for instance a balance which includes subjective weight on post-treatment quality of life (Krumholz et al. 2006a). Additionally, AMI diagnoses have an incidence rate and mortality rate which are high enough to allow for statistical inferences. Common methods of estimating patient risk use administrative data on patient morbidities upon arrival. Administrative data is quite detailed and estimating patient risk from administrative data can require a reduction in the dimensionality of patient morbidity detail. Methods to reduce this dimensionality include selecting a subset of highly predictive admission morbidity codes (Pine 2007) and categorizing different morbidities together based on clinical coherence, such as the hierarchical condition classification used by the Centers for Medicare & Medicaid Services, (CMS; HHS 2011). The quality of a prediction is measured by the precision and accuracy of an estimate and by a statistic called the area under the ROC curve or c-statistic. Variation in accuracy within a condition is minimal, and so simulating variation in precision (and thus in the c-statistic) is the main focus of introduced noise in the simulations. While having slightly different scales, estimates of both of these methods vary from between 0.9 and 0.6, with precision 1.0 being an implausibly perfect prediction.

To parameterize the simulation all inpatients from the New York State Inpatient Databases 2005-2007 whose primary diagnosis is AMI (ICD-9-cm codes starting with 410, excluding those with an indicator for a prior occurrence) are pooled. The New York State Inpatient Databases collected by the Healthcare Costs and Utilization Project contains approximately 90% of the state's inpatient discharges each year (HCUP 2014). The data collected is primarily used to calculate reimbursements for patient care. Variables include the patient's age, sex and a set of patient comorbidities encoded using of ICD-9-CM diagnosis codes. The first code listed is designated as the primary diagnosis.

The sample is limited to cases between 2005 and 2007 in which Acute Myocardial Infarction (AMI) is either the primary or a secondary diagnosis. The age and sex of the patient, the source of admission (self-admit, emergency, and transfer are the main categories), as well as the primary and eight secondary diagnoses are used as predictors. Age is binned into five year intervals. Along with those, this paper uses dummy variables for the patient sex and five dummies for different patient admission source along with 6,286 3 digit ICD-9-CM based binary comorbidity indicators totaling 6337 predictors, all of which are binary variables. Using binned age indicator variables is commonly recommended as a simple means of controlling for non-linear age effects. Tree based models such as random forests control for the non-linear effect of age on mortality implicitly, and using a single continuous variable or multiple indicator variables has no effect on the results. Indicator variables are used for comparability to other research.

Patient in-hospital mortality risk is estimated from these variables including only comorbidities recorded as present upon admission. The simulation adds noise to the patient risk variable to simulate variation in quality of patient risk estimation. To estimate patient risk as a single variable, first patient mortality probability is estimated using a random forest following the method of the second chapter which uses all administrative data on patient morbidities coded in ICD-9 as well as ages and hospital admission source to model patient mortality. This allows us to estimate with high precision and accuracy patient mortality probabilities, which is used as base estimates of patient risk.

In such a large data set, population mortality rate (0.0951) is nearly exactly equal to the mean patient risk, 0.0951 (sd for patient risk 0.151). The number of AMI patients seen at each hospital ($npat_h$) varies from 1 to 1363 (mean = 165, sd = 188). Following the method used by CMS (Grady et al. 2013) observations from hospitals that saw less than 25 patients are dropped, leaving 52,194 patients and 324 hospitals. This step reduces a number of hospitals that perform very poorly or perfectly, small simulations which include these hospitals gave results consistent with those presented below. Hospital risk-adjusted mortality rates are estimated from the entire set of observations in the five unique ways available from the set in Table 3.1: “Raw local standardized”, “Local fixed effects”, “Sample fixed effects”, “Local random effects”, and “Sample random effects”. When estimated for the entire population, “Fitted sample standardized” equals “Sample random effects,”, and “Fitted local standardized” equals “Local random effects.” Formulae for each RAMR calculation are given in Table 3.2, discussed below.

Table 3.1: Summary Statistics

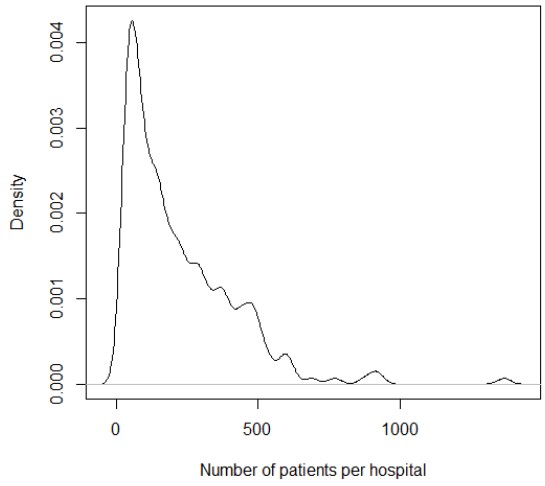
Variable	Mean	SD	Min	Max
Patient Risk	0.0955	0.152	0.0394	0.992
Patients per Hospital	220	190	25	1363
Hospital Mortality Rate	0.111	0.0509	0	0.389
RAMR ("Raw local standardized")	0.102	0.0389	0	0.304
RAMR ("Fitted local standardized")	0.104	0.0367	0.0092	0.289
RAMR ("Local fixed effects")	0.111	0.0409	<0.001	0.389
RAMR ("Sample fixed effects")	0.103	0.0513	<0.001	0.359
RAMR ("Local random effects")	0.114	0.0513	0.0102	0.392
RAMR ("Sample random effects")	0.106	0.0410	0.0160	0.363

Note: Patient risk is summarized over all patients, while the other variables are at the hospital level. . "Raw" refers to an estimate which does not use a model of patient mortality. "Local" refers to a calculation which only uses patients which are actually admitted to each hospital while "sample" uses modeled patient outcomes if all patients were admitted to each hospital. "Standardized" refers to the standardization of an estimate by multiplying modeled outcomes by the ratio of actual mortality to expected mortality. "Fixed" effects use a fixed indicator variable for a hospital in modelling patient outcomes while "random" effects fit hospital outcomes in the model to a normal distribution with an estimated mean and standard error, and then use the mean of the distribution as a point estimate of the hospital-specific effect. Hospitals with less than 25 patients are dropped. Patient risk is fitted estimates from a logistic model of mortality on patient morbidities. RAMRs are estimated from pooled NY and CA data, 2005-2007.

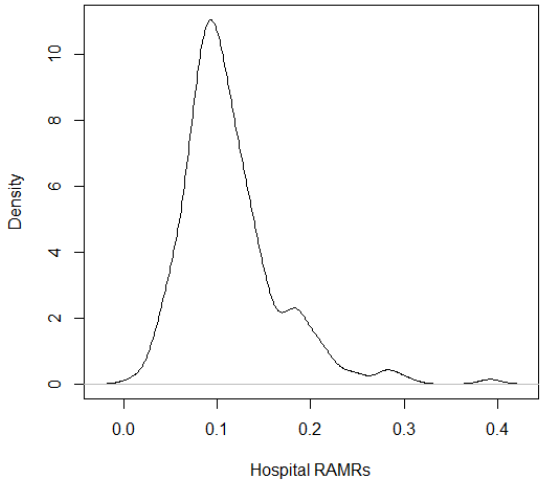
The different estimates of RAMR have nearly identical means and standard deviations. While not perfect matches, the “fitted local standardized” method used by CMS (Grady et al. 2013) is selected as the basis for the simulations. Figure 3.1 presents a KDE distribution of the number of patients per hospital and of average RAMRs estimated from patient scores. Both variables are truncated at zero and have long right tails. Since hospitals with less than 25 patients are excluded, average RAMRs are neither too close to zero nor too close to one.

Figure 3.1: Distributions of key variables, NY and CA AMI inpatients, 2005-2007

Panel 1: KDE of number of patients per hospital.



Panel 2: KDE of average hospital RAMRs from patient scores.



Note: For completeness, all hospitals are included in Panel 1. Hospitals with less than 25 patients are dropped in panel 2 and in all subsequent analyses. Estimates are from pooled NY and CA data, 2005-2007.

Hospitals do not receive patients randomly from the universe of patients. That is, the mean patient risk of a hospital's patients may be correlated to the RAMR of a hospital. The nature of this relationship is complex; it will be related to the decision of a patient's ambulance team, the relationship between a patient and a hospital, the socio-economic condition of the patient's neighborhood and socio-economic aspects of patient risk not captured in the patient's age and morbidities, etc. (see Finlayson et al. 1999). Likewise, hospital RAMR may be correlated to the number of patients a hospital receives, N_h, either through patients preferring hospitals with lower RAMR or through hospital learning by doing. The purpose of this paper is not to understand the subtleties of these factors (for more, see Pitches et al. 2007), but it is important to test if these factors affect the quality or RAMR formulae.

3.4. Model

3.4.1. Estimating Risk-adjusted Mortality Rate

There are various possible formulae that can be used to estimate risk-adjusted mortality rate. Before discussing these, this section needs to briefly define a few variables. Let $patientrisk_{hi}$ be an exogenous, unbiased measure of a patient's risk of in-hospital mortality, $mort_{hi}$. Average mortality and average patient risk for a particular hospital are given by $\overline{mort}|_h$ and $\overline{patientrisk}|_h$, while averages for the entire sample are $\overline{mort}$ and $\overline{patientrisk}$. In this analysis the sample will be considered to be the entire population of cases. In cases where the sample does not equal the population, a further distinction would need to be made in these averages. This analysis relies on two different risk models with hospital effects. The fixed effects model is estimated as:

(1) $mort_{hi} = invlogit(\beta_{fe} * patientrisk_{hi} + \alpha_{fe,h} + \epsilon_{fe,hi})$,

Where $\epsilon_{fe,hi} \sim N(0, \sigma_{fe})$ is a disturbance parameter and $\alpha_{fe,h}$ is a hospital-specific effect estimated as a fixed effect in the model. Note that there is no global intercept, so a hospital-specific effect is estimated for every hospital. The expected mortality at a given hospital for patients of that hospital from this model is denoted $E_{i\in h}[invlogit(\beta_{fe} * patientrisk_{hi} + \alpha_{fe,h})]$. The expected mortality at a given hospital for patients of that hospital from this model is denoted $E_{\forall i}[invlogit(\beta_{fe} * patientrisk_{hi} + \alpha_{fe,h})]$.

A less common alternative formation for hospital RAMR use is $invlogit(\beta_{fe} * E_{i\in h}[patientrisk_{hi}] + \alpha_{fe,h})$ and $invlogit(\beta_{fe} * E_{\forall i}[patientrisk_{hi}] + \alpha_{fe,h})$. While the expectation of the inverse logit of a variable does not equal the inverse logit of its expectation, these two estimates are very similar and this alternative is not included. Note that the hospital specific effect is generally an underestimate

of RAMR unless $patientrisk$ were normalized to have its mean equal zero. In that case the inverse logit of the hospital specific effect is equal to this alternative formulation, i.e. $invlogit(\alpha_{fe,h}) = invlogit(\beta_{fe} * E_{i\in h}[patientrisk_{hi}] + \alpha_{fe,h})$.

The random effects model is estimated as:

(2) $mort_{hi} = invlogit(\beta_{re} * patientrisk_{hi} + (\alpha_{re,h} + \epsilon_{re,hi}))$,

Where $\alpha_{r,eh} \sim N(\mu_{re}, \delta_{re})$ is estimated as normally distributed from a probability model and $\epsilon_{re,hi} \sim N(0, \sigma_{re})$ is the second disturbance parameter. Again there is no global intercept, so it is not necessary to assume that $E(\mu_{re}) = 0$. Otherwise, this specification matches what is normally called random effects in the economics literature. A fitted estimate of the probability of mortality for an observation from this model is denoted $\widehat{mort_{hi}}|_{re}$. The expected mortality at a given hospital for patients of the hospital from this model is denoted $E_{i\in h}[invlogit(\beta_{re} * patientrisk_{hi} + \alpha_{re,h})]$. The expected mortality at a given hospital for patients of that hospital from this model is denoted $E_{\forall i}[invlogit(\beta_{re} * patientrisk_{hi} + \alpha_{re,h})]$. As with "fixed effects", the average of the set of mortality risks for the set of patients is the focus rather than estimating hospital mortality risk for the average of the set of patients.

This paper suggests seven ways to estimate risk-adjusted mortality rates which can be characterized across a number of different strategy groupings. One strategy grouping involves estimating the ratio of the overall population mortality rate divided by an estimated expected mortality ratio and multiplying this ratio by the hospitals actual or predicted mortality ratio. The paper calls this grouping "standardized" estimation, the ratio of population mortality to hospital or sample mortality can be considered a standardization of risk-adjusted mortality provided by the expected mortality rates. Occasionally risk-standardized mortality rates (RSMR) is used in the literature instead of risk-adjusted mortality rate. A distinction between the two is not always made, but in this paper, standardized refers only to mortality rates estimated in this way.

A second grouping involves estimating hospital-specific effects in a logistic regression of patient outcome on patient risk. If the hospital-specific effects are estimated from a probability model to follow a normally distributed hospital-specific component of the error in the logistic regression, this is the method commonly called random effects in the economics literature, and this paper will denote this group "random effect" estimation. Alternatively, hospital-specific fixed effects can be included in the logistic regression, to be denoted "fixed effect" estimation.

Another dimension of grouping is to decide if hospital risk-adjusted mortality rates are to be estimated for the patients the hospital actually receives or if they are to be estimated as if the hospital has received an average patient from the sample (or, nearly equivalently, the hospital has received the entire set of patients in the sample). This paper uses the terms “local” and “sample” to differentiate between these. Finally, “local standardized” estimation of hospital risk-adjusted mortality rates can be calculated two ways, a hospitals actual mortality ratio can be divided by the expected mortality ratio or a hospitals mortality ratio can be predicted using a model which includes hospital effects and that prediction can be divided by expected mortality ratio. The former this paper calls “raw local standardized” estimation and the latter this paper calls “fitted local standardized” estimation. The latter method is performed using a hospital random effects model to predict patient mortality for hospitals in the method used by CMS (Grady et al. 2013), and this paper follows that procedure. Racz & Sedransk (2010) give the name of Bayesian risk-adjusted mortality rate to a version of “fitted local standardized” estimation of RAMR where the average patient risk in the denominator is estimated using Bayes rule. This paper does not include this method, but a Bayesian estimate of the average patient would be nearly identical to a frequentist estimate in the case used in this simulation where the set of predictors is very limited and very small hospitals are dropped. Table 3.2 gives a precise description of each method.

Table 3.2: Risk-adjusted Mortality Rate Estimation Formulae

Estimation name	Formula	Use in Literature
"Raw local standardized"	$\frac{\overline{mort} * \overline{mort}\|_h}{\overline{patientrisk}\|_h}$	Guru et al. 2008; Hannan et al. 2013;
"Fitted local standardized"	$\frac{\overline{mort} * E_{i\in h}[invlogit(\beta_{re} * patientrisk_{hi} + \alpha_{re,h})]}{\overline{patientrisk}\|_h}$	Krumholz et al. 2006b; Grudy et al. 2013
"Fitted sample standardized"*	$\frac{\overline{mort} * E_{\forall i}[invlogit(\beta_{re} * patientrisk_{hi} + \alpha_{re,h})]}{\overline{patientrisk}}$	-
"Fitted fixed effects sample standardized"*	$\frac{\overline{mort} * E_{\forall i}[invlogit(\beta_{fe} * patientrisk_{hi} + \alpha_{re,h})]}{\overline{patientrisk}}$	-
"Local fixed effects"	$E_{i\in h}[invlogit(\beta_{fe} * patientrisk_{hi} + \alpha_{fe,h})]$	Mark et al. 2005
"Sample fixed effects"*	$E_{\forall i}[invlogit(\beta_{fe} * patientrisk_{hi} + \alpha_{fe,h})]$	-
"Local random effects"	$E_{i\in h}[invlogit(\beta_{re} * patientrisk_{hi} + \alpha_{re,h})]$	-
"Sample random effects"*	$E_{\forall i}[invlogit(\beta_{re} * patientrisk_{hi} + \alpha_{re,h})]$	-

Note: A selection of recent uses of a formula in the literature is included. "Raw" refers to an estimate which does not use a model of patient mortality. "Local" refers to a calculation which only uses patients which are actually admitted to each hospital while "sample" uses modeled patient outcomes if all patients were admitted to each hospital. "Standardized" refers to the standardization of an estimate by multiplying modeled outcomes by the ratio of actual mortality to expected mortality. "Fixed" effects use a fixed indicator variable for a hospital in modelling patient outcomes while "random" effects fit hospital outcomes in the model to a normal distribution with an estimated mean and standard error, and then use the mean of the distribution as a point estimate of the hospital-specific effect.

** Standardizing "Local fixed effects" is identical to "Raw local standardized". Because this paper uses a sample as a population, "Fitted fixed effects sample standardized" is the same as "Sample random effects" and "Fitted sample standardized" is the same as "Sample random effects" in all models in this paper.*

The taxonomy has three pieces, “local” vs “sample”, “raw” vs “fitted”, and among the fitted, “fixed” vs “random”. This suggests a pattern, and at first glance the pattern is broken in two places. First, “Raw sample standardized” is not included because “raw” refers to actual and not modeled outcomes and “sample” refers to all admits in a sample and not just to those who are admitted to a particular hospital, but actual patient outcomes can only be observed at the hospital to which they are admitted. Second, If patient risk is estimated from an unbiased model and the sample is the entire population, then $E(\overline{patientrisk}) = \overline{mort}$, and “Fitted sample standardized” estimation is equal to “sample random effect” estimation. In this analysis, the patient risk model is not forced to be unbiased (since this paper uses a nonlinear estimate and do not allow observed patient risk < 0), but in every case the difference between $\overline{patientrisk}$ and $\overline{mort}$ is nearly undetectable and so only results for “sample random effect” estimation are reported.

If the expected mortality rate at every hospital is equal to the expected mortality rate at all hospitals, which would be the case if patients were randomly distributed to hospitals, then $E(\overline{patientrisk}|_h) = \overline{mort}$. In this case, “fitted local standardized” estimation would be equal to “local random effect” estimation and “raw local standardized” estimation would be equal to “local fixed effect” estimation.

However, even if *patientrisk* is unbiased, if patients prefer to attend higher quality hospitals, i.e. $cor(\overline{patientrisk}|_h, N_h) > 0$, assuming patient severity is not correlated with hospital quality $cor(\overline{patientrisk}|_h, RAMR_h) = 0$, then more patients will attend hospitals with lower expected mortality rates and the expected average patient risk at each hospital will be less than the overall mortality rate, $E(\overline{patientrisk}|_h) < \overline{mort}$. If patient severity is correlated with hospital quality, the direction of this inequality may change, but the underlying point is that an estimate of “fitted local standardized” need not equal an estimate of “local random effect” and an estimate of “raw local standardized” need not equal an estimate of “local fixed effect”. In this analysis, these estimates are not equal so both estimates are reported.

Before moving on, consider three notes about theoretical considerations. First, the loss functions used in optimizing likelihood functions over a non-linear function are not amenable to estimating a theoretical comparison of the goodness-of-fit estimates, and rather, simulation and numerical estimation are appropriate. Second, random effects are commonly avoided in economics because they require special conditions to consistently estimate coefficients on the fixed effect. In this case, the fixed effect is patient risk and the estimate of interest is the hospital-specific effect. This fact limits the applicability of this concern to this analysis, although fitted RAMR does use the coefficient on patient risk in its estimation. Once could, thus, interpret the use of random effects estimates in this analysis as allowing that a theoretically inconsistent estimate of the fixed effect combined with a more

efficient estimate of the hospital-specific effect may ultimately be more efficient and useful than other methods. Third, it is also important to note that focusing on "raw local standardized" estimation is not intuitively satisfying for a number of reasons. One, the ratio of observed to expected mortality for a high risk, high mortality facility and for a low risk, low mortality facility need not consistently compare the quality of the two facilities. Two, while it may be argued that using a model of hospital mortality means the hospital is being judged on theoretic rather than actual patient outcomes, a hospital's treatment of its patients plays a role in those patients' estimated risk scores used in the denominator of this (and all) models, and thus including theoretic outcomes of a hospital's patients cannot be avoided.

3.4.1. Simulation

Since performance of these formula cannot be readily compared theoretically, this paper develops a simulation which allows the formula to be compared with a true RAMR many times with many different amounts of noise. The first task of the simulation is to aggregate a distribution of patient risks and mortalities to estimate hospital RAMRs. $patientrisk_{hi}$ is taken from a kernel density estimate (KDE) of patient risks from the data and draw $mort_{hi}$ from a binomial distribution using the patient risks.[8] The number of hospitals is set equal to the number of hospitals in the data. Each hospital is assigned a "true" RAMR ($rawRAMR_h$) for each simulation, estimated from a KDE of RAMR from the "Local random effects" estimate.

In the simulation, observations are assigned a hospital by a bootstrap method. Each patient's hospital is selected randomly with replacement from list of hospitals equal in length to the number of patients, but with hospitals allowed to repeat so they appear in the same the frequency they appear in the data. Thus the distribution of patients per hospital will match the actual distribution, but there will be some variation between the observed number in any simulation and the expected number.

To add interesting variation in RAMR and hospital case mixes with plausible parameters, two simple OLS regressions are estimated. The first regression allow inclusion of a component of hospital quality that is endogenous to the simulation, that it is correlated with the number of patients the hospital treats.

[8] Using a KDE ensures that using a specific set of patient risk scores does not obscure uncertainty in patient risks. An alternative to drawing from a KDE of patient risk estimates estimated in the previous chapter is to estimate a set of patient risks using the methods of the previous chapter but using a bootstrapped sample of the data. Using a KDE instead of bootstrapping patient risk scores could provide greater variation between simulations as the bootstrapped sample will, on average, have more identical observations than draws from the KDE.

(3) $\widehat{RAMR_h} = \alpha_{(3)} + \beta_{(3)} * npat_h + \epsilon_h$,

An additional benefit of including this relationship in the model is that it will later be able to use fitted RAMRs to estimate $\beta_{(3)}$ and thereby test the degree of attenuation bias from measurement error in different estimates of RAMR.

The second relationship tested considers the relationship between average patient risk and RAMR.

(4) $\overline{patientrisk}|_h = \alpha_{(4)} + \beta_{(4)} * \widehat{RAMR_h} + \epsilon_h$

In both regressions, the number of observations is equal to the number of hospitals (258). The estimated value from (3) of $\alpha_{(3)} = 0.14$ $(sd = 0.0057)$ and of $\beta_{(3)} = -0.00011$ $(sd = 0.000015)$ and from (4) of $\alpha_{(4)} = 0.066$ $(sd = 0.0038)$ and of $\beta_{(4)} = 0.339$ $(sd = 0.030)$. Parameters are set to conservative (closer to zero) round numbers: $\alpha_{(3)} = -\frac{1}{10000} * \overline{npat_h}$; $\beta_{(3)} = \frac{1}{10000}$; $\alpha_{(4)} = -\frac{1}{5} * \overline{RAMR_h}$; and $\beta_{(4)} = \frac{1}{5}$.

In order for hospital size variation to be simulated as in (3), this simulation include a hospital size effect to be added to RAMR:

(5) $sizeeff_h = -\frac{1}{10000} * npat_h + \frac{1}{10000} * \overline{npat_h}$

In this estimation, the number of patients per hospital from the data is used, rather than the bootstrapped patient-hospital assignment. From this the hospital RAMR is calculated:

(6) $RAMR_h = rawRAMR_h + sizeeff_h$

In order for patient selection variation to be simulated as in (4), a hospital-specific selection effect is added to each patient's patient risk:

(7) $selecteff_{ih} = \frac{1}{5} * rawRAMR_h - \frac{1}{5} * \overline{rawRAMR_h}$

Note that letting $\alpha_{(3)} = \beta_{(3)} * \overline{npat_h}$ and $\alpha_{(4)} = \beta_{(4)} * \overline{RAMR_h}$ results in:

(8) $E(sizeeff_h) = E(selecteff_{ih}) = 0$

Thus these effects do not change the means of the patient risk and RAMR estimates that they modify.

The model specific and hospital-specific nuisance parameters are combined into a single nuisance variable which is normally distributed with mean zero and standard deviation having a simulation-specific component as well as a component specific to each hospital within a simulation. In a given

simulation, all patient risks are measured with a noise, having a variance, $\rho \sim N\left(0, \left(\frac{1}{10}\right)^2\right)$. This noise simulates measurements with different levels of precision. Within each hospital in each simulation the noise is increased by adding to ρ, a hospital-specific noise component given by $\nu_h \sim N(0, \nu^2)$, which allows different hospitals to have different levels of precision. The degree of hospital-specific noise varies across simulations according to $\nu \sim N\left(0, \left(\frac{1}{10}\right)^2\right)$. The nuisance for any given observation will then be a mean-zero random variable with normal distribution and variance equal to the sum of ρ, the total noise on the patient risk for a particular run of the simulation, and ν_h, the noise on patient risk for patients at a particular hospital in that run of the simulation: $\eta_{ih} \sim N(0, \rho^2 + \nu_h^2)$.

Observed patient risk in the simulation is then:

(9) $patientrisk_{ih} = rawpatientrisk_{ih} + selecteff_{ih} + \eta_{ih}$

Mortality is simulated from "true" patient risk without the noise parameter but including the selection effect, $mort_{hi} \sim binom(patientrisk_{hi} - \eta_{ih})$ – so $P(mort_{hi}) = patientrisk_{hi} - \eta_{ih}$. Observed and "true" patient risks are censored to be in the set, $[0, 1]$, RAMR censoring was unnecessary as no RAMRs were outside of $(0, 1)$.

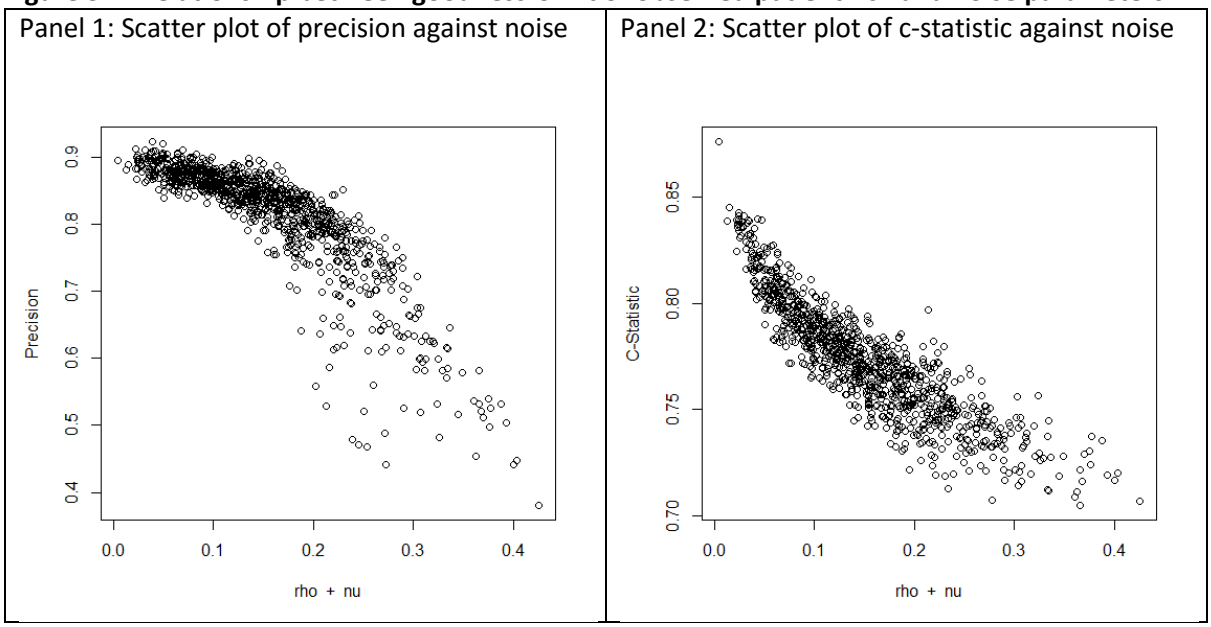
The simulation is repeated 1,000 times and in each simulation RAMR is estimated six different ways according to models (1) and (2) and the formulae in Table 3.2 using $patientrisk_{hi}$, $mort_{hi}$, and the patient's simulated hospital assignment. Also equation (3) is estimated from every fitted $\widehat{RAMR}$, and every observed $\widehat{npat}$.

3.5. Results

Comparisons between formulae are based on their performance across four measures as the amount of noise in the simulated *patientrisk* varies. For each observation of *patientrisk* and patient mortality this paper also calculates the precision and the c-statistic as correlates to noise that may be estimated in real world data. In order to motivate the relationship between noise and observed patient risk, this paper first present scatter plots of precision against the sum of noise parameters, $\rho + \nu$, and of the c-statistic against $\rho + \nu$ (Figure 3.2). These plots show the model to cover the universe of goodness of fits common in the risk-adjustment literature (see second chapter, Pine 2007). Since ρ and ν are independent, simulations where one is low and the other is high are likely. The sum, $\rho + \nu$, is a measure of the total amount of noise in a simulated trial. When noise is very low, the precision and c-

statistics are close to 1; while as the noise parameters increase, these decrease to lower values similar to those found when risk-adjustment is performed using less granular data.

Figure 3.2: Relationship between goodness-of-fit of observed patient risk and noise parameters



Note: Given the rule that an observation is expected to die if observed patient risk is above 50%, precision equals the number true positives divided by the total number of positives. The c-statistic is equal to the probability that a randomly chosen positive instance will have higher observed patient risk than a randomly chosen negative one. Estimates are from pooled NY and CA data, 2005-2007.

Quality of RAMR estimation is judged by comparing simulated $RAMR_h$ and each fitted $\widehat{RAMR_h}$ for the six models. The four measures used are:

$$RMSE = \sqrt{mean((\widehat{RAMR_h} - RAMR_h)^2)}$$

$$Bias = mean(\widehat{RAMR_h} - RAMR_h)$$

$$SpearmanCorrelation = cor(\widehat{RAMR_h}, RAMR_h)$$

$$AttenuationBias = \widehat{\beta_{(3)}} - \widetilde{\beta_{(3)}}$$

Where $\widetilde{\beta_{(3)}}$ comes from an estimate of model (3) using simulated $RAMR_h$ and simulated $\widehat{npat}$ and represents the relationship that would be estimated if true RAMRs were known, while $\widehat{\beta_{(3)}}$ is estimated from estimated $\widehat{RAMR_h}$ and simulated $\widehat{npat}$. Using simulated $\widehat{npat}$ for both estimates of $\beta_{(3)}$ allows the estimate of attenuation bias to be more applicable to estimates of $\beta_{(3)}$ than would be made in the literature where there is no "true" number of patients per hospital other than the "observed" number of patients per hospital. Analysis that uses $\beta_3 = \frac{1}{10000}$ gives similar results to using $\widetilde{\beta_{(3)}}$. Comparisons are presented in Figure 3.3 across variation in noise on patient risk, ρ and in Figure 3.4 across variation in heterogeneous, hospital-specific noise on patient risk, ν. Here this paper pays close attention to "raw local standardized" estimation and "fitted local standardized" estimation, as these are the most common methods. A clear way to present trends in the goodness-of-fit measures, this paper presents the results using local scatterplot smoothing (loess) of the plot of the measure against noise parameters. Loess gives a smooth curve through this data and is appropriate since the shape of the relationship between these measures and noise parameters is not known. 95% confidence intervals of the curve are presented as dashed lines around the loess estimate. The estimates are made using a smoother span of 0.75 with 2 degree polynomials.

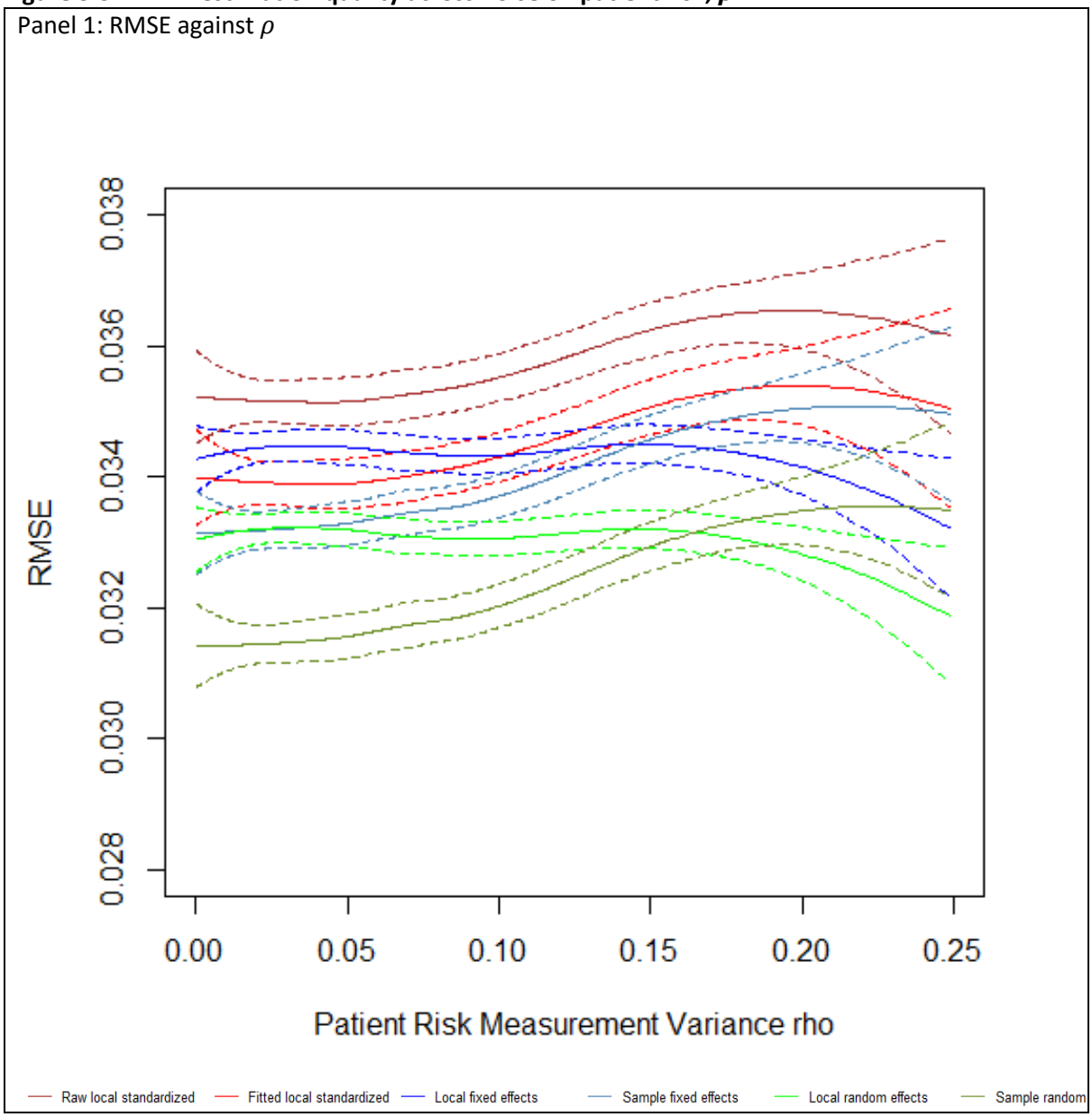
Minimizing SSE is the optimization condition of ordinary least squares regression. This condition is equivalent to minimizing RMSE, but unlike SSE, RMSE has the same scale as the data, and thus gives us a good idea how wrong, on average, an estimated RAMR is. In this case, "raw local standardized" and "fitted local standardized" estimates are two of the worst three, while "sample random effects" and "local random effects" are the best and second best. Because the difference between these is only the standardization, this suggests that standardization does not improve model fit. Bias estimates the direction, on average, of the error in estimation of RAMR. In this case the two "fixed effects" estimates perform worst, underestimating RAMR by about one percentage point (or about 10% of mean RAMR). RMSE is a better measure of how incorrect an estimate of RAMR is, but accuracy in both

of these measures is important for estimation of the difference in health outcomes when choosing between hospitals.

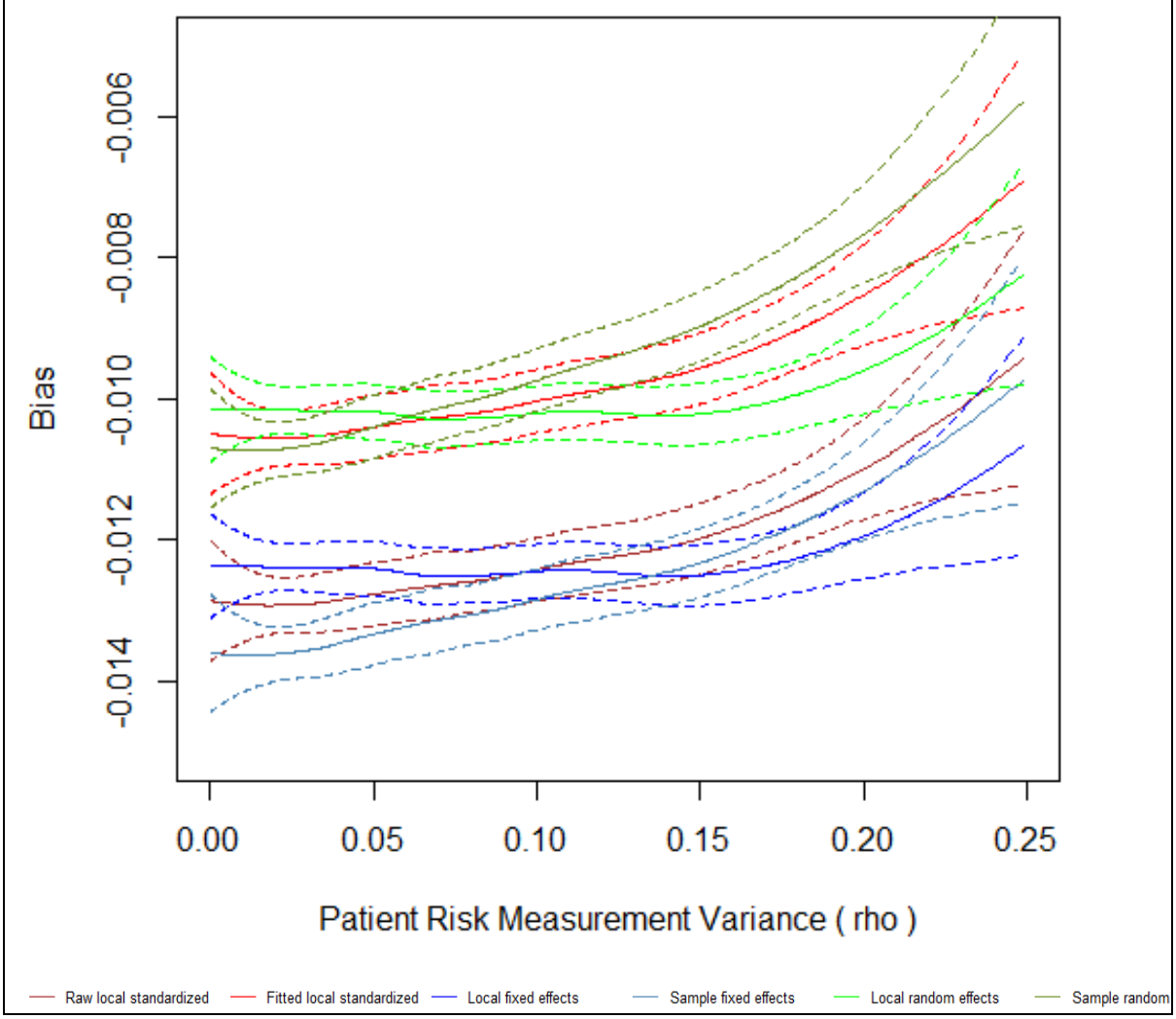
Correlation of estimated RAMR to actual RAMR is extremely important for the credibility of a pay-for-performance program that emphasizes relative position of ranking of hospitals. Low correlation implies that many hospitals will be incorrectly judged to be performing below or above average. In this case again "local random effects" performs best and the two "direct" estimates perform worst.

The final estimate, bias in estimation of $\beta_{(3)}$ provides a flavor of how important accurate estimation of RAMR is for scholars seeking to estimate the determinants of hospital quality. All bias estimates suggest a negative bias. Since the "true" $\beta_{(3)}$ is negative, this means that the estimations of $\beta_{(3)}$ using estimated RAMRs is larger (and closer to zero), consistent with interpreting poor estimation of RAMR as measurement error and this bias as attenuation bias. Since the true value of $\beta_{(3)} = 10^{-4}$, the mean bias of $1.2 * 10^{-5}$ is an error of 12%, so the scale of this bias is quite important and will play a role in the ability of a researcher to identify a signal and properly recognize its sign. For this statistic, "fitted local standardized" estimation does perform best, and "local random effects" performs second best. "Raw local standardized" is in the middle of the pack, suggesting that scholars should avoid using this measure when possible.

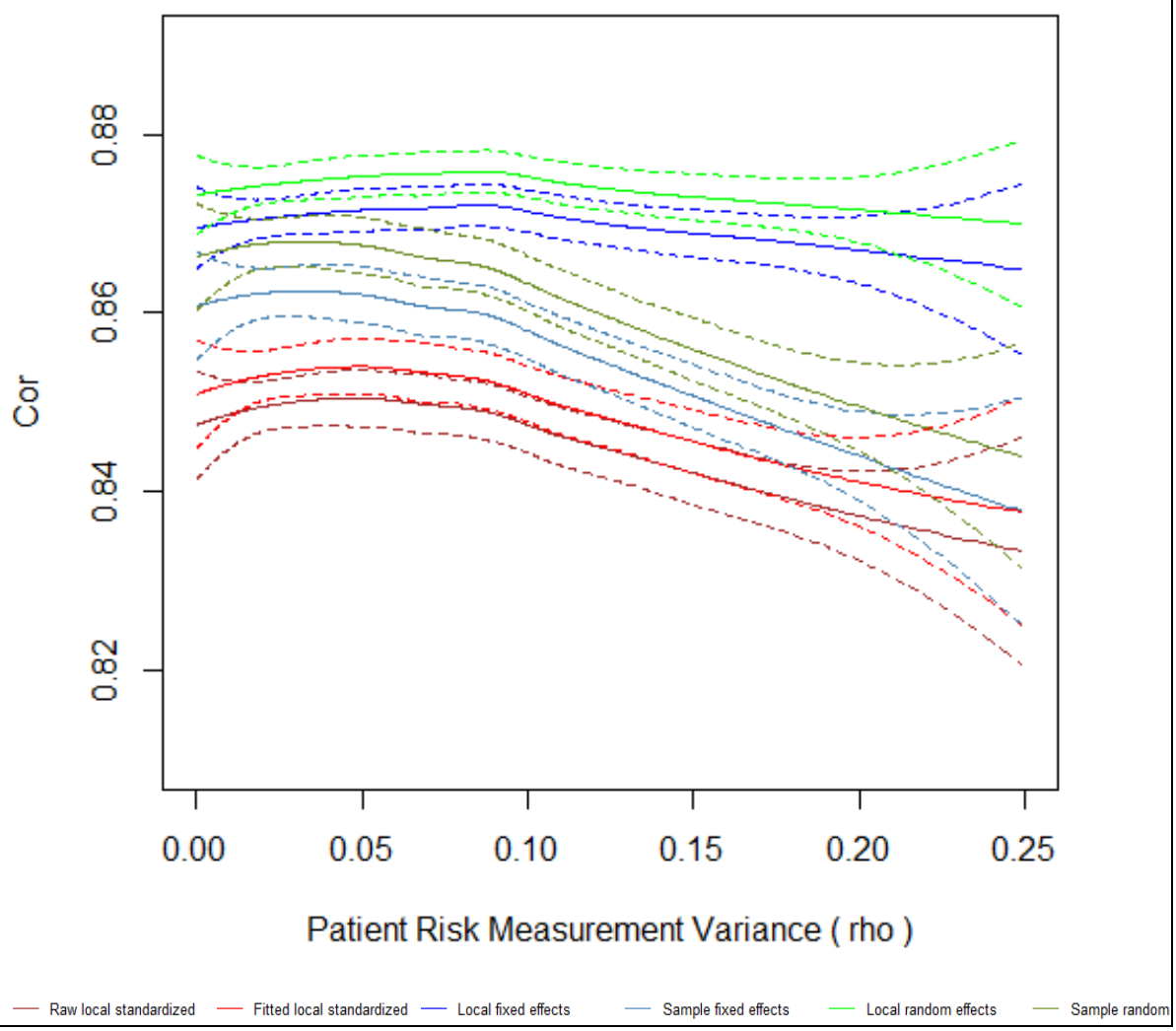
Figure 3.3: RAMR estimation quality across noise on patient risk, ρ



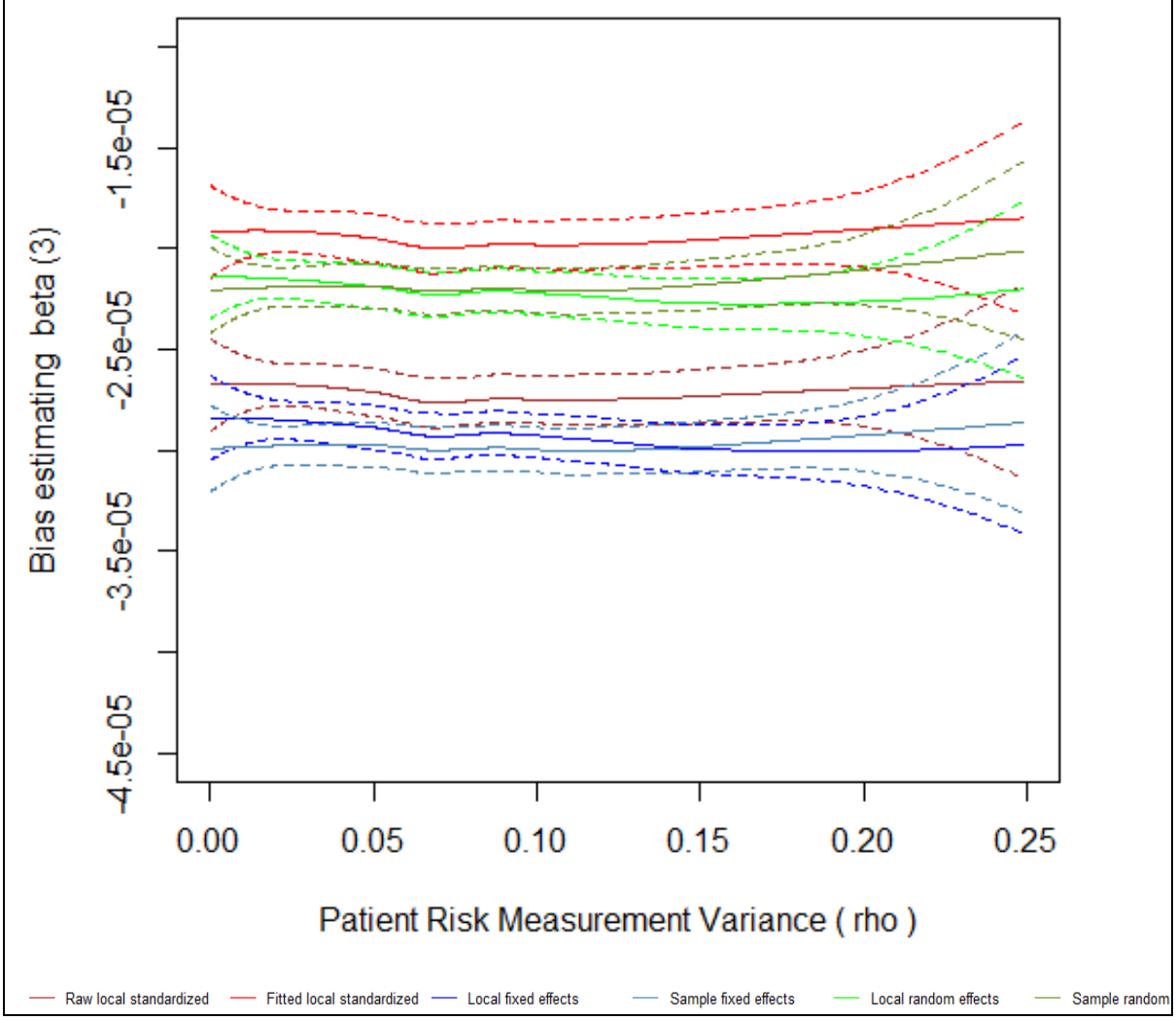
Panel 2: RAMR Bias against ρ



Panel 3: Correlation against ρ

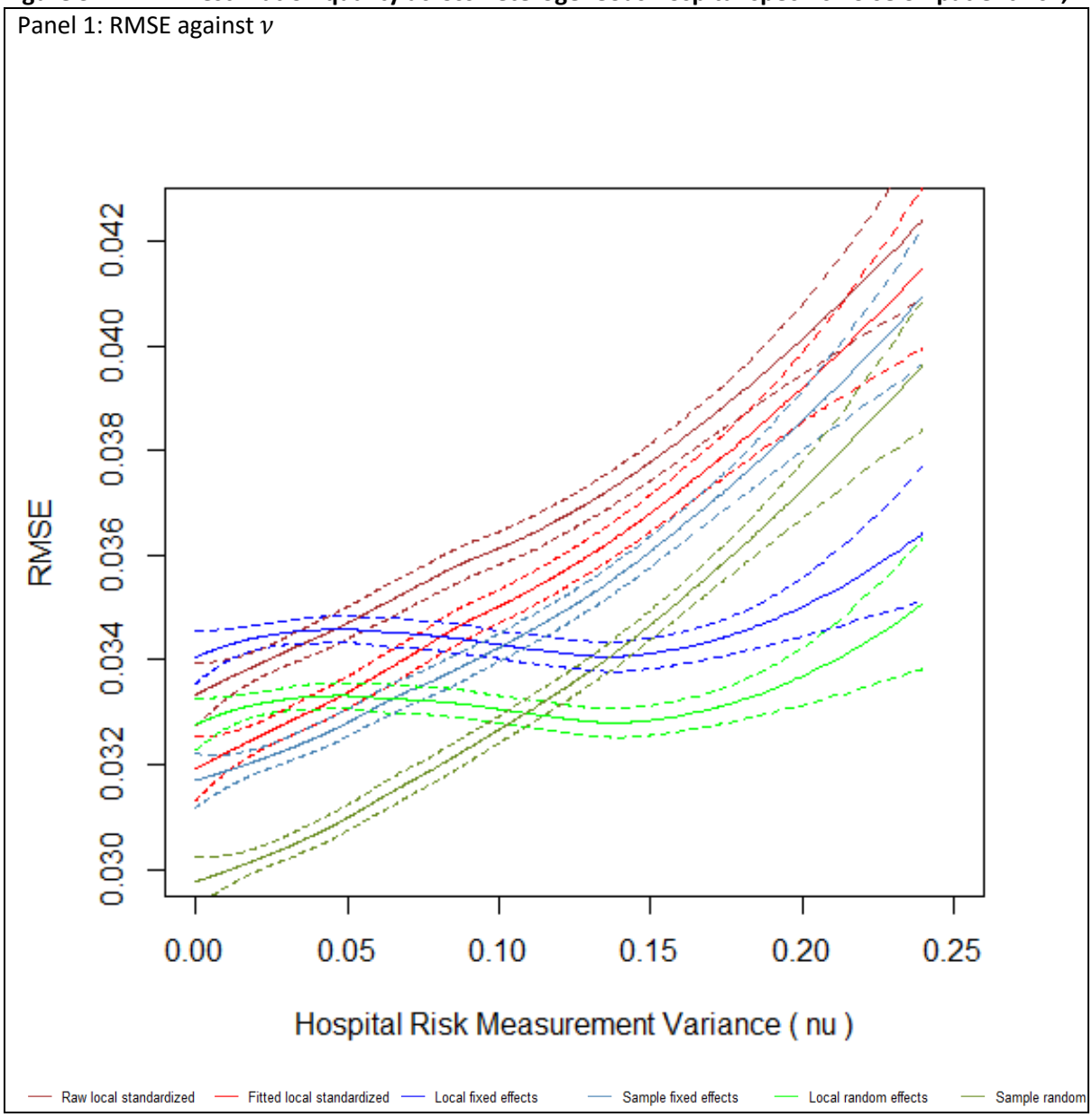


Panel 4: Bias estimating $\beta_{(3)}$ against ρ

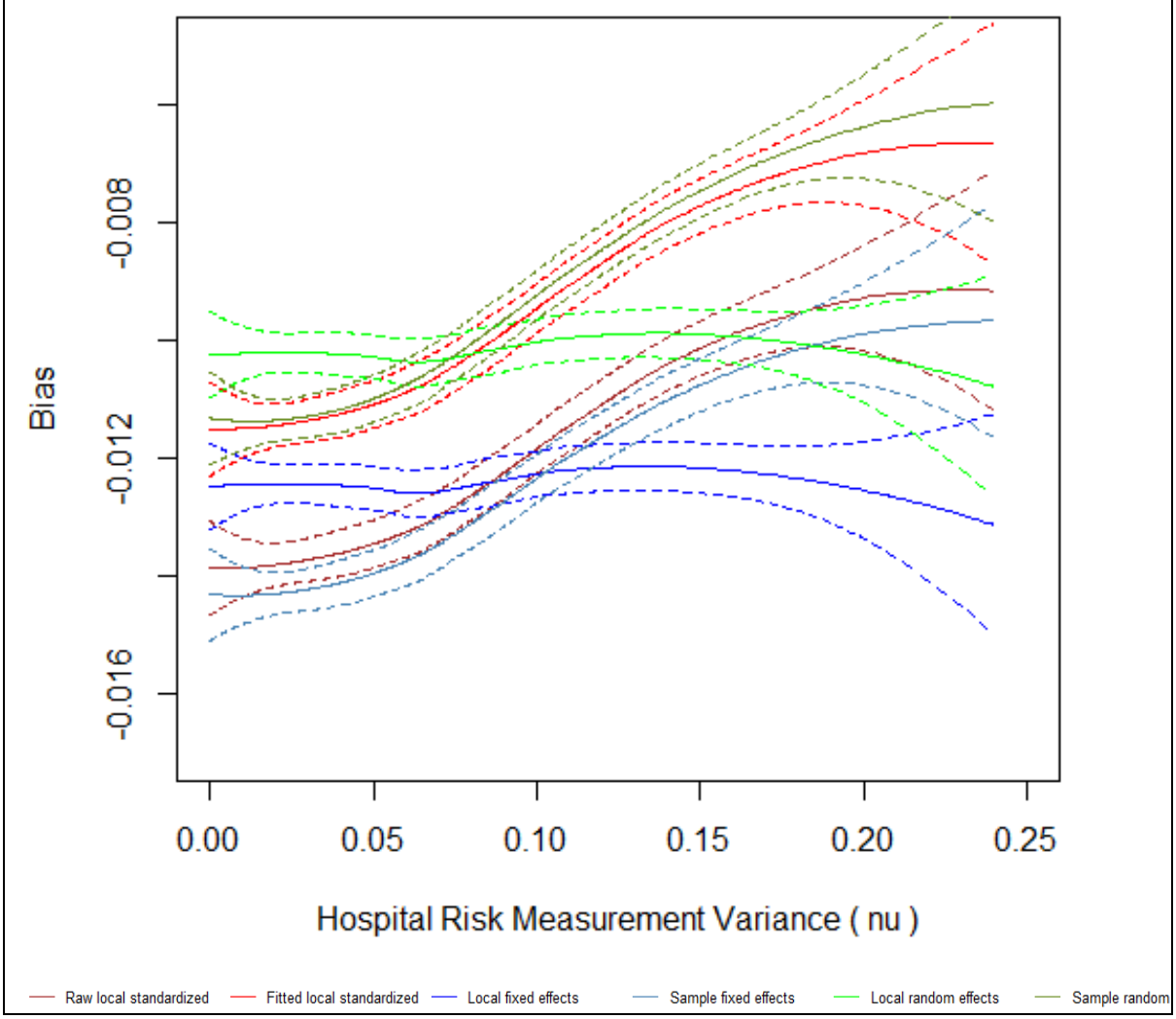


Note: Four measures of goodness of RAMR fit as a function of simulated patient risk measurement variance fitted using loess curves with dashed lines representing 95% confidence intervals, data from 1,000 simulations built from New York State Inpatient Databases 2005-2007 AMI inpatients.

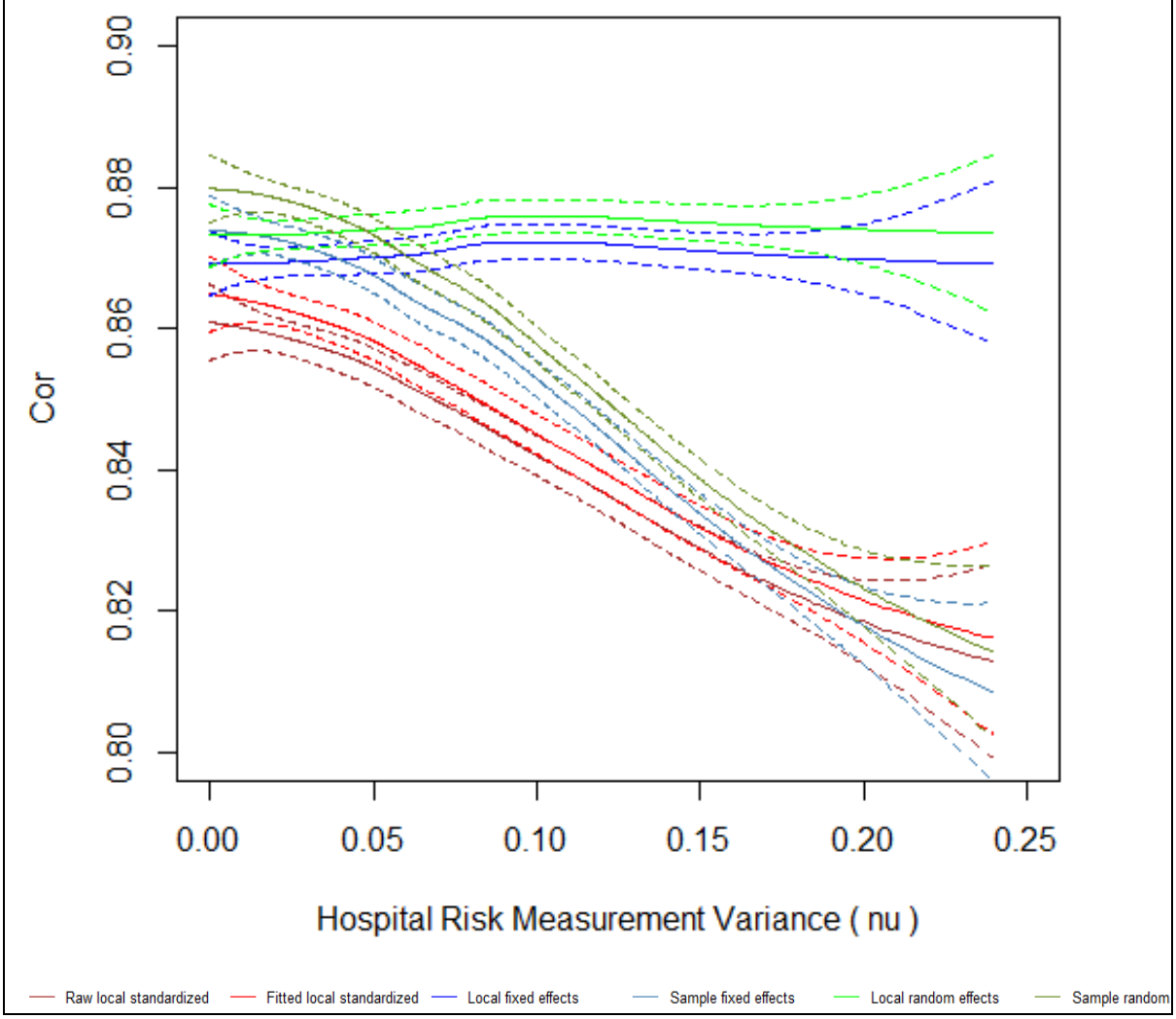
Figure 3.4: RAMR estimation quality across heterogeneous hospital-specific noise on patient risk, ν



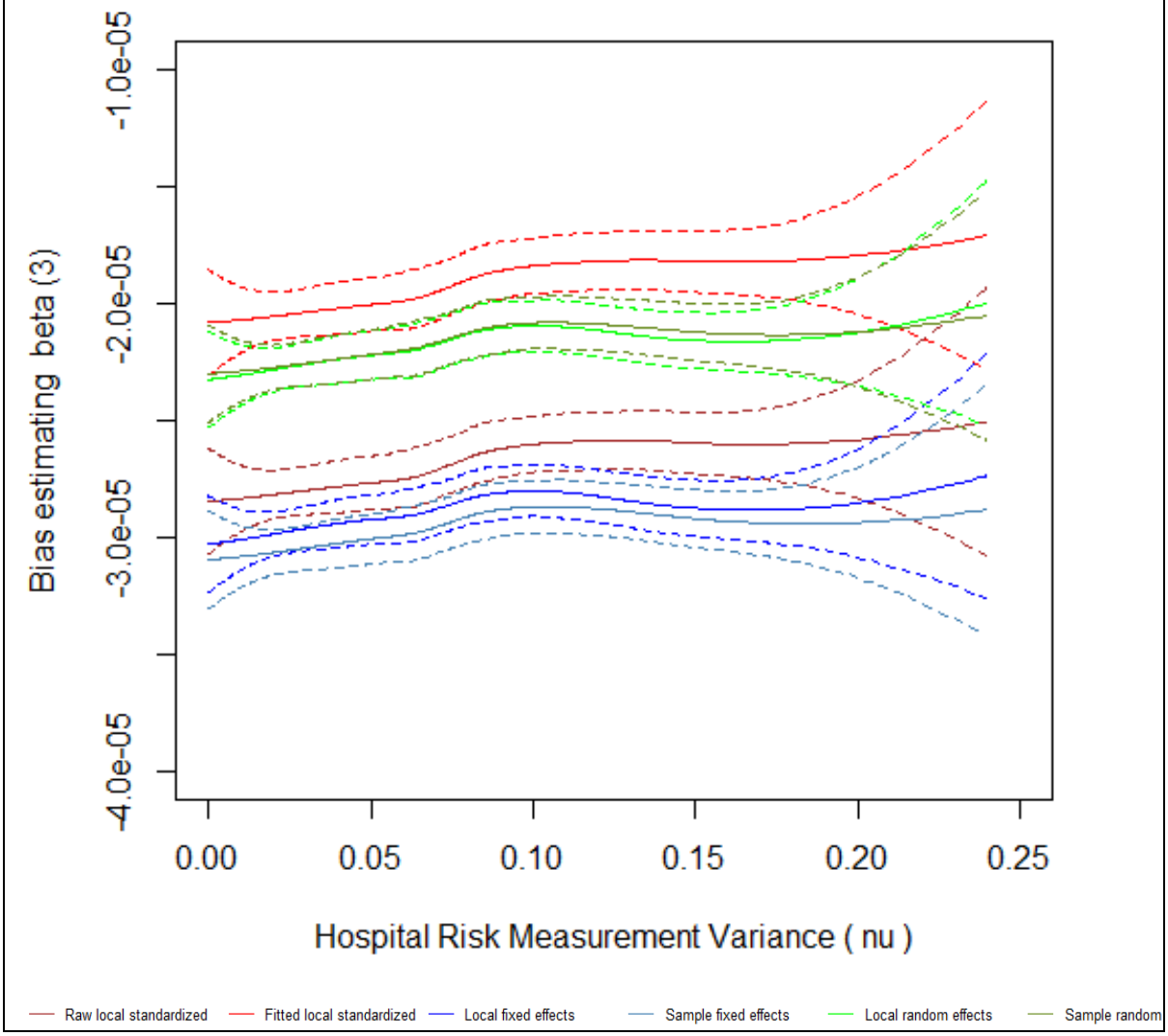
Panel 2: RAMR Bias against ν



Panel 3: Correlation against ν



Panel 4: Bias estimating $\beta_{(3)}$ against ν



Note: Four measures of goodness of RAMR fit as a function of simulated hospital risk measurement variance fitted using loess curves with dashed lines representing 95% confidence intervals, data from 1,000 simulations built from New York State Inpatient Databases 2005-2007 AMI inpatients.

Hospital-specific levels of noise can simulate a context where patient risk precision varies across hospitals. This simulates non-uniform hospital reporting of patient morbidity data. The general trends in this dimension are similar to those across the patient risk measurement noise. However, there is an important difference in how different measures of RAMR respond to increased hospital-specific noise. "Local fixed effects" and "local random effects" do not greatly vary as hospital-specific noise increases in any measure, but bias, RMSA, and especially correlation are very sensitive to the degree of hospital-specific noise. Therefore it is particularly important to pay attention to the RAMR measure used in contexts with inconsistent reporting of patient risk across hospitals.

In the estimate of attenuation bias in the volume-quality relationship, in cases with nonzero values of nu, ν, the measurement error in hospital quality is endogenous in the simulation. This model did not include potential instruments, neither instruments which deal with volume-quality endogeneity, nor instruments which deal with heterogeneous measurement error in quality estimation. However, this endogeneity has no effect on the applicability of the result –these recommendations apply to research involving volume and quality as well as to research involving non-endogenous determinants of quality. This paper did perform tests with nu, ν, set equal to zero to simulate an exogenous determinant of hospital RAMR. In these tests, the relationship between RAMR estimation quality and noise on patient risk, ρ, show a nearly identical profile of attenuation bias to those seen in Figure 3.3 regardless of RAMR formula.

It is important to remember that the estimation of the volume-quality relationship presented in this chapter is at the aggregate, hospital level, while the real outcome of interest, patient mortality, is a patient level variable. The true relationship between hospital quality, hospital volume, and patient mortality should be estimated using a multilevel model. The linear presentation of a simple relationship between hospital volume and hospital RAMR presented in this paper is similar to that found in many places in the literature, but can only give an aggregate effect estimate, and cannot be interpreted as the effect of volume on the probability of an individual patient's death.

3.6. Conclusion

Risk-adjusted mortality rates (RAMRs) are an important and commonly used measure of hospital quality. A number of different formulae are used to generate RAMRs from patient risk estimates with little concern given to the role these formula play in affecting the usefulness of the estimated RAMRs. This paper shows that RAMRs estimated using common techniques are not optimal for a number of important purposes and that non-standardized versions should be considered. Particularly, standardizing estimates reduces their usefulness for four tasks health economists typically use in their

work. Standardizing a hospital mortality rate is necessary when the goal is to use a hospital's actual mortality rate to estimate its RAMR, but, a hospital's predicted mortality rate is often preferred to its actual rate (Grady et al. 2013). When the hospital's mortality is predicted using hospital-specific effects, standardizing does not further improve the estimate. Rather, when a model with hospital-specific effects is used to predict in-hospital mortality, that prediction is itself a preferable estimate of RAMR. RAMR estimates when hospital effects are fitted with a probability model (such as in random effects models) allowing shrinkage of estimates towards the mean further improves the estimation compared to estimating hospital effects as fixed effects.

This paper presents a number of formula for estimating RAMR, the most common two being "raw local standardized" or "fitted local standardized". In many cases, these results recommend using a different formula. The biggest impact of the results may be in pay-for-performance estimation since this scheme is most efficient when it correctly ranks hospital quality. This paper finds that the Spearman correlation of simulated true RAMR to estimated RAMR using the methods commonly used by policy makers to perform significantly worse than other methods. Depending on the scheme, this difference will play a significant role in how well a scheme aligns hospital interests to its performance. If a hospital judges that a scheme cannot discern quality signal from noise, it may reduce the willingness of a hospital to prioritize investment in quality improvement. For similar reasons, this paper also find that patient choice should not be based on the most popular methods of calculating RAMR.

Measurement error in RAMR when using RAMR as an outcome in regression models did result in attenuation bias, and that this bias is great when using what this paper calls "raw local standardized" estimates of RAMR, a method that is common. In this case this paper recommends "local random effects" or "fitted local standardized". The magnitude of the bias ranges from over 16% of the signal, while for high levels of noise, using "local random effects" can reduce bias to less than 6% of the signal. This difference has significant implications for hypothesis testing and estimating effect sizes in modeling hospital quality.

The aggregation of interest in this work is hospitals, but the results could be applicable to any institutional or jurisdictional grouping, such as the RAMR for a particular geographic region or facility type, or even, plausibly, for coherent groupings of patients, such as when comparing RAMR across states. These results may also apply to risk-adjusted rates of hospital readmission, which is another important measure of hospital quality.

Chapter 4

4.1. Introduction

Provision of health care provides an interesting opportunity for economists to study the relationship between volume and quality. This paper focuses on provision of care for inpatients admitted with Acute Myocardial Infarction (AMI), which is a condition frequently used in research on hospital quality. Quality of provision is measured using risk-adjusted mortality rates which is a measure of a hospital's mortality rate which takes into account non-random case mixes in hospitals. There is a negative relationship between most kinds of hospital volume, including AMI volume, and RAMR; as volume increases, quality improves resulting in a decrease in RAMR. This relationship may be in part due to patient selection, patients prefer better care, increasing volume at better hospitals. The relationship may also be in part due to returns to scale and/or learning-by-doing, that is there may be a causal effect of volume on quality. This paper's contribution is to use a novel instrument to estimate that causal effect.

Instrumental variables are frequently used to address endogeneity in the volume quality relationship in health provision. A common instrument is the total volume of patients with similar conditions within a certain radius of each hospital. This paper addresses critiques of that instrument by looking at the volume of another set of conditions a hospital sees, the volume of shock and of trauma patients. This instrument is correlated with AMI volume, but since it may not be correlated with patient choice, an estimate using this instrument can provide evidence for causality in what otherwise is an endogenous relationship. Using this instrument, this paper does find evidence for causality, even in cases where hospital-specific fixed effects are included – in contrast with recent research (Kim et al. 2016).

The relationship between risk-adjusted mortality rates and patient volume is often measured using cross-sectional data, an approach which exploits variation between hospitals. An advantage of panel data is that repeated measurements of a hospital allow a model to be estimated using hospital-specific effects. Controlling for hospital-specific fixed effects results in an estimate of the relationship which is measured within a given hospital. In this way, fixed-effects control for time invariant hospital factors, the most important of which for this analysis is the size of the hospital infrastructure. Including hospital fixed effects is overlooked in some existing research, and a recent paper (Kim et al. 2016) argues that when hospital-specific fixed effects are included, the relationship is no longer significant. This paper is consistent with that result in that including hospital-specific effects in the model greatly

increases the standard error around the estimate, but does find some evidence that a within hospital effect of volume on quality exists.

In addition to using hospital fixed effects this work introduces the novel use of shock and trauma volume as an instrument for AMI volume to deal with endogeneity in the volume-quality relationship. Emergency technicians responding to patients who are in shock or have experienced trauma might prioritize minimizing the time getting the patient to the hospital as quickly as possible. If they do, then trauma and shock volume reflect the volume of patients near a hospital. Thus, this sort of volume may be correlated with the volume of AMI patients, but is not otherwise correlated with the quality of the hospital treatment of AMI cases. This instrument is generally strong but greatly weakened when hospital-specific fixed effects are included.

This paper starts with a discussion of the relationship between volume and quality and the current state of research. It then presents three models commonly used to estimate this relationship, two of which use as their unit of analysis the hospital, and one of which is at the individual level. The following section discusses the data used in this paper. Then, in the results section, estimates are given for the main model including some evidence for a causal relationship, even when hospital-specific fixed effects are included. The robustness section considers some variations, especially in how volume is measured, and how the relationship is modeled. The paper then concludes with a discussion of its limitations and policy implications.

4.2. Background

The association between hospital volume and mortality is an active area of research, with most research finding that higher volume is associated with better outcomes. In fact, the relationship is so robust, Silber et al. (2010) suggest including volume as a predictor of quality to improve the accurate estimation of hospital risk-adjusted mortality rates. More common, however, is the study of the relationship between these two variables and its implications.

One important application of this research is to identify cut points and inflections in the relationship between volume and quality. Ross et al. (2010) identify a level at which returns to volume taper off. Their study is based on about 3.5 million Medicare patients with AMI, heart failure, or pneumonia in the United States from 2004-2006. Another group of studies considers the other side of the quality-volume relationship, suggesting that at very low volumes, quality tends to be poorer, and that this implies that certain procedures should be avoided where possible at low volume facilities. Gutacker

(2016) studies the relationship between mortality in coronary artery bypass graft surgery (CABG) and volume of those surgeries. The study includes hospitals in five countries, Denmark, England, Portugal, Slovenia, and Spain. He finds that mortality is much lower in large, English hospitals and conclude that hospitals which would have annual volume less than 415 refrain from the surgery. His focus is common in CABG volume research, an area where the debate frequently focuses on minimum safe volumes, ranging from 450 to 150 and even lower (Shahian 2004).

Another task in this area of research is to understand the mechanism behind the relationship. Schull et al. (2006) study the relationship between hospital volume and accurate myocardial infarction (MI) diagnoses. They consider the role of experience in MI care, suggesting that volume could be a proxy for both experience and the presence of better skilled specialists. Their work suggests better standardization of care to reduce the difference in measured quality between high- and low-volume hospitals.

Not all work in this area finds a significant quality-volume relationship. Lee et al. (2015) consider ischemic heart disease (IHD) cases in Victoria, Australia between 1998 and 2005 and do not find a significant relationship between volume and quality. However, when controlling for overall volume, they do find that specialization has a positive relationship with quality. To measure specialization, they estimate the number of IHD cases as a proportion of all cases the hospital sees. While this paper does not look at specialization, the suggestion of no-effect in Lee et al. (2015) and Kim et al. (2016) must be considered carefully. Particularly, Kim et al. (2016) shows the importance of using hospital-specific effects for understanding the policy implications of the result. Since many policies relating to volume are unlikely to change the hospital specific effect, it is important to include these effects when seeking to use estimates to make policy recommendations.

An important limitation in papers including these is that endogeneity in the volume-quality relationship may result in estimates which are not causal. There are two competing explanations for the volume-outcome relationship. One is economies of scale, possibly through learning-by-doing. The other is selective referral or patients being more likely to attend higher quality hospitals, either through their own selection or through the selection of a referring agent, for instance by an ambulance dispatcher or another physician of the patient. There are important policy implications related to which of these mechanisms is driving the volume-outcome relationship. If volume improves quality through economies of scale or learning by doing, then consolidation of patients to centralized hospitals may improve overall public health. On the other hand, if patient selective referral drives this result, then optimal policy may hinge on whether there is a crowding effect from this selection, or if

some economies of scale persist. A number of papers use instrumental variables to attempt to deal with this endogeneity.

In a paper studying the relationship between CABG mortality and volume, Gaynor et al. (2005) attempt to estimate a causal relationship between volume and quality, explicitly as a way to measure returns to scale and learning-by-doing. As an instrument for CABG volume, they use the number of CABG patients residing in and the number of CABG-offering hospitals operating in various fixed geographical radii around the given hospital. This instrument is found to be very strong, suggesting that the number of nearby procedures and hospitals are good predictors of a hospital's volume. However, they decide not to include the instrument in the main analysis on account of their Hausman-Wu test statistic being above 0.05. In fact, their Hausman-Wu test statistic has a very low p-value, 0.06. Even so, they do not reject the null hypothesis of exogeneity and instead argue the instrument is unnecessary. They argue that if quality affects volume through selection, an instrumental variables estimate of the volume effect would be lower than the non-instrumental variables estimate of the volume effect, while they found an increase in the coefficient on volume. Using a probit regression of individual patient mortality on hospital volume, their results predict that at the mean volume level in their data (216 cases per year), an increase in volume by one case would reduce mortality by 0.003 percentage points. An increase in volume across the interquartile range, from 98 to 263 cases, would reduce mortality by 1.38 percentage points from 2.39% to 1.78%.

Another paper by Gowrisankaran et al. (2006) uses a model where three kinds of mortality including CABG mortality are estimated using a logit function at the patient level as a function of hospital volume and again uses the number of patients within a specific distance to the hospital as instruments. Their estimates imply that increasing quarterly volume from 117 CABG cases to 175 cases reduces the chance of dying by 0.5 percentage points from 3.6% to 3.1%.

As with the non-causal papers, there are some examples where no significant relationship is found. Heusch (2009) uses a similar instrument to Gaynor and to Gowrisankaran to estimate learning-by-doing effects for CABG patients of individual surgeons rather than for hospitals. Their econometric set up uses a patient-level model in the second stage. This paper uses quarterly data for 57 Florida surgeons during the 36 quarters between 1998 and 2006. This paper also sets up a model of forgetting, where surgeon experience can depreciate. Almost all of their reported estimates are very noisy, and in their instrumental variables models their t-statistics are uniformly smaller than one in absolute magnitude, and they reject the hypothesis of a causal volume-quality relationship. Another paper by Ramanarayanan (2008) uses the occurrence of a surgeon stopping performing CABG as an exogenous shock to increase volume for other surgeons at a facility as an instrument again for the volume-

outcome relationship in CABG cases. While this paper finds a significant effect, it is much smaller than other research; an increase of one additional case per year decreases mortality by 0.005 percentage points the following year.

Instead of using cases in an area around a hospital as an instrument for hospital volume, this paper uses volume of shock and trauma patients at the hospital as instruments for hospital AMI volume, as patients with these conditions will be less likely to be allowed to spend more time in an ambulance going to a preferred hospital, and instead will be sent to the nearest possible hospital. The importance of shock and trauma in this paper is based on the idea that emergency technicians responding to patients who are in shock or have experienced trauma might prioritize minimizing travel time to the hospital. If they do, then trauma and shock volume reflect the volume of patients near a hospital. Thus, this sort of volume may be correlated with the volume of AMI patients, but is not otherwise correlated with the quality of the hospital treatment of AMI cases.

Scale or volume is sometimes measured as a cumulative variable, but recently this has been criticized, and in any case it is usually measured contemporaneously with patient outcomes. In fact, Gowrisankaran et al. (2006) and Heusch (2009) both include forgetting in their models; compared to factories with a small number of outputs doctors face a large variety of patients and comorbidities and the value of experience gained may be lost when new patient's with different illnesses arrive. Lee et al. (2015) argue that learning is more important in the context of specialization; while many physician's skills are fungible, this suggests that forgetting is less likely or that learning is more successful when variation in patients is reduced.

Much of the research uses ordinary least squares to study the relationship between a hospitals risk-adjusted mortality rate and its patient volume. Lee et al. (2015) note that when risk-adjusted mortality rates are calculated in a separate estimation, this two-step method may induce heteroscedastic errors and standard errors should be calculated using the Huber-White sandwich estimator. It is also common to perform estimation at the patient level, looking at individual mortality as a function of hospital volume using a probit or logit model. Kim et al. (2016) look at in-hospital mortality of Florida, New York, and New Jersey cancer patients between 2000 and 2011 who are given surgical procedures for colectomy, esophagectomy, pancreatic resection, pneumonectomy, pulmonary lobectomy, or rectal resection. Their work estimates patient mortality on annual hospital volume and finds that significance of their results depends on whether or not hospital fixed effects are included, results are not significant when annual hospital fixed effects are included. They note that in their fixed effects specification the estimation requires variation in annual hospital volume within a hospital, which may

be limited and greatly reduce the power of their estimate, but suggest that there remains significant variation in annual volume within a hospital and that this issue does not drive their null result.

The instrument used in this paper assumes that emergency medical services (EMS) prioritize minimizing the time between their loading the patient into an ambulance and the patient's arrival at the hospital when patients are diagnosed with shock and/or trauma. In emergency medicine, the "golden hour" refers to findings that the patient outcomes deteriorate when a patient with severe injury waits more than an hour for care. It has long been recommended that minimizing total out-of-hospital time is particularly important in trauma and shock patients (Poitras 2011). The importance of this period has been challenged for trauma victims in recent studies, although it likely holds for patients who are in shock. Newgard et al. (2015) consider time between an incident and arrival at the hospital, or out-of-hospital time, for patients with shock and patients with traumatic brain injury. They find no evidence that lower out-of-hospital time improves outcomes in traumatic brain injury patients, and mixed evidence that such time improves outcomes in patients with shock.

McCoy et al. (2013) separate out-of-hospital time into response time (the time it takes EMS to arrive at the scene), scene time (the time EMS spends at the scene), and transport time (the time it takes once the patient is loaded in the ambulance to arrive at a hospital) in trauma patients. They did not have data on response time, but they do compare the importance of scene time and transport time. They find that scene times longer than 20 minutes may worsen outcomes in some trauma patients (particularly penetrating trauma), but do not find transport times play a significant role in outcomes. Swaroop et al. (2013) and Crandall et al. (2013) both use the Illinois Trauma Registry and find a similar result that minimizing out-of-hospital time for penetrating trauma victims is critical.

Emergency care research often focuses on cardiac arrest, heart conditions, trauma, and motor vehicle accidents. However, some research considers all patients. Pons et al. (2005) look at EMS response times for all patients and find mixed evidence that they are critical for survival, a very short response time (less than four minutes) did improve outcomes in their data, but when time was modeled continuously they did not find a significant relationship between time to response and outcomes. Wilde (2013) also looks at the relationship between EMS response time and mortality and find that mortality rates worsen as time to care increases and also looks at all patients. Their study uses distance between the incident and the nearest EMS facility as an instrument. This instrument is meant to deal with endogeneity that arises when the emergency service knows patient severity and reduces response time when patient severity is high.

Newgard et al. (2010) add another category to out-of-hospital time, activation, which they define as the time between onset and contacting emergency services. This gives five time variables: activation, response, scene, transport, and total time in a study of trauma patients. This study does not find a significant effect of variation in any of these time variables on patient outcomes. Fleet and Poitras (2011) point out limitations in Newgard's work and suggest that lives are saved when out-of-hospital time is minimized. Harmsen et al. (2015) include most of these and other less recent papers in a systematic review of the relationship between out-of-hospital time and outcomes in trauma patients. Their review concludes that out-of-hospital time is important, although there is some evidence that increased on-scene time may improve outcomes. It concludes, however, that minimizing transport time is important as well.

This paper, then, seeks to add to the literature first by introducing a novel instrument, Volume of shock and trauma patients. Second, this paper follows a number of important methodological suggestions in the literature: using robust standard errors, including hospital-specific fixed effects, instrumenting hospital volume, including both hospital-level and individual-level models, and discussing scale versus learning.

4.3. Method

4.3.1. Instrument Validity

The instrument used in this paper is the volume of patients with trauma and or shock during that month. Patients who experience trauma or shock need to be stabilized as quickly as possible. When a patient has these conditions, they will be taken to the facility which can accept them which involves the shortest delay possible. This suggests the use of the volume of trauma or shock patients a hospital receives as an instrument for monthly AMI volume.

Instrument validity is based on the satisfaction of two primary assumptions, the "exclusion restriction" that the instrument has no effect on the outcome except through the predictor, and that the instrument is associated with the predictor being instrumented. The second assumption is tested using an F-test which estimates how strongly the instrument predicts the endogenous predictor controlling for covariates. The first assumption cannot be directly tested, and is based on the design.

One common instrument for patient volume is the number of cases within a certain distance of a hospital. Both the number of nearby cases and the number of trauma and shock cases are functions of the number of people near the hospital who could possibly have an AMI, and thus are related to

the number of ill people who could go to the hospital as instruments for the number of people who do. This factor, the size of the population the hospital could serve, could influence the level of investment a hospital receives, but this investment may be unlikely to change over the period studied and will be captured in the fixed effect. Instead of using cases in an area around a hospital as an instrument for hospital volume, this paper uses volume of shock and trauma patients at the hospital as instruments for hospital AMI volume, as patients with these conditions will be less likely to be allowed to spend more time in an ambulance going to a preferred hospital, and instead will be sent to the nearest possible hospital. One advantage of trauma and shock cases as an instrument over nearby cases is that the definition of the radius may affect the instrument. In cases within a fixed distance around the hospital is used, variations in population density and road density within that ring can influence the demographics within that ring and quality of the hospital. Less dense population could mean that a hospital's catchment is significantly larger than the ring used in the instrument, and so for these hospitals the fitted volume from the first stage of the IV will be underestimated. If hospitals in low density areas are systematically of different quality than hospitals in more dense areas, this will bias the estimate – for instance if rural hospitals are worse than urban hospitals, using a radius-based instrument may overestimate the effect of volume on quality, making low volume hospitals look worse than they actually are, as volume at low quality rural hospitals is underestimated in the first stage. This instrument can be modified to handle population density and road structure around the hospital, but such modifications are ad hoc at best, as these variables are not constant within a ring around the hospital. Another criticism of using cases within a certain distance of a hospital is that population home location choice may be partially determined by proximity to a good hospital. Using trauma and shock volume may be preferred in this case as well, victims of trauma or shock may be less likely to predict the possibility of their future illness when choosing a location of their homes and thus may not be influenced by the quality of the nearest hospital.

Medical conditions in the hospital administrative data in this paper are categorized using codes from the International Classification of Disease Volume 9 – Clinical Modification (ICD-9-CM or ICD-9).[9] There are numerous different kinds of shock and of trauma. In order to simplify the analysis, four codes for shock are considered. These codes are grouped in that they start with the same three-number ICD-9 code, 785: unspecified shock (ICD-9 code 785.50), cardiogenic shock – shock resulting from failure of the heart to pump an adequate amount of blood as a result of heart disease and especially heart attack (ICD-9 code 785.51), septic shock – a life-threatening form of sepsis that usually results from the

[9] ICD-9-CM was the scheme used in the US until October, 2013 and is used in this data. Since October, 2013, Volume 10 of the ICD codes have been used in the US. ICD-10 is not perfectly resolvable into ICD-9 codes, but is nearly so and results in this work should not depend on the coding scheme used.

presence of bacteria and their toxins in the bloodstream and is characterized especially by low blood-pressure and reduced blood flow to organs and tissues and often organ dysfunction (ICD-9 code 785.52), and other (ICD-9 code 785.59). Volume of patients within each of these categories is considered separately and then all cases are considered together, creating 5 shock volume variables. Similarly, ten types of trauma based on ICD-9 code 958 are considered: air embolism as an early complication of trauma – an air embolism is an obstruction of circulation by air in the veins usually introduced by wounds (958.0), fat embolism as an early complication of trauma – in this case the circulatory embolism is by fatty tissue (958.1), secondary and recurrent hemorrhage as an early complication of trauma (958.2), posttraumatic wound infection not elsewhere classified (958.3), traumatic shock (958.4), traumatic anuria – a condition sometimes known as crush syndrome and associated with trauma to the kidney (958.5), Volkmann's ischemic contracture – a condition caused by obstruction of the brachial artery near the elbow (958.6), traumatic subcutaneous emphysema – or gas trapped under the skin, often caused by a punctured lung (958.7), other early complications of trauma (958.8), and traumatic compartment syndrome – a condition arising from internal bleeding into a compartment of the body containing nerves and muscles which cannot easily stretch to contain it and can lead to a loss of blood flow to the area (958.9). All types are grouped into a single volume variable.

The importance of shock and trauma in this paper is based on the idea that emergency technicians responding to patients who are in shock or have experienced trauma might prioritize minimizing travel time to the hospital. If they do, then trauma and shock volume reflect the volume of patients near a hospital. Thus, this sort of volume may be correlated with the volume of AMI patients, but is not otherwise correlated with the quality of the hospital treatment of AMI cases.

Another issues to address when considering instrument validity is the assumption that a given observation is unaffected by treatments assigned or received by other observations, an assumption called the stable unite treatment value assumption. This assumption is frequently violated in cases like this study where geographic variation plays a role and geographic spillovers may occur. This violation may cause bias of uncertain direction. In order to control for this issue, results are clustered by geographic area, although this can only partially control for the issue, as geographic areas in the data may not exactly match the spillover areas which lead to the bias (Green and Vavreck 2008).

4.3.2. Model

This paper's main model will study patient outcomes at an aggregate level. For robustness, it will also look at a model of the individual level. At the aggregate, hospital level, a hospital's mortality rate is a

function of a number of patient-level and hospital-level variables. Most of the patient-level variables are captured in a measure of patient risk. For this reason, including patient risk is a key component of any model of hospital quality. One important patient level variable that is only partially captured in the data available in this paper relates to the amount of time it takes for the patient to receive care. The amount of time before the patient is admitted to the hospital will be partially captured in patient severity data. The amount of time the patient spends admitted to the hospital but not being properly cared for will be partially captured by hospital fixed effects. Two other hospital level factors are important as well, the hospital's human capital level and physical capital level, which represent skill and training in the hospital and infrastructure in the hospital. Over short time periods, these two factors are more or less fixed for a hospital. One important aspect of hospital infrastructure is the size of the hospital. The size of the hospital, as might be measured by the number of beds in the hospital or the full-time equivalent staff levels is also largely fixed over a short time period.

Fundamental intuition applying the economics of scale to hospital care would be based in the ability of large hospitals to invest in specific, specialized tools and staff. This investment plays a key role in the intuition behind scale cut-offs, volume below which it is not recommended that certain procedures be carried out, such as the CABG study by Gutacker et al. (2016). This type of specialization will, also, generally be fixed in a hospital over a short time period. The remaining relationship between exogenous volume and patient outcomes may, then, be ascribed to learning and practice by the hospital's physicians, nurses, and staff.

Various measures of hospital volume are possible. Cumulative hospital volume could be a discounted or non-discounted count of a hospital's patients since some starting point. Given the possible role of forgetting and the importance of recent volume, this paper looks at current-month volume and at past-month volume.

Patient severity is estimated using the model described in Chapter 2. Using the set of patient comorbidities information directly, this paper creates a very large set of predictors. Sparse, high-dimensional models often do not converge and reducing dimensionality means losing information and leads to the potential for consequential omitted variable bias. A patient's mortality likelihood is estimated using a random forest with actual in-hospital mortality as the outcome variable and these comorbidity indicators as well as patient age, sex, and mode of arrival at the facility as predictors. The fitted patient score from this model is regressed in a logit model against actual outcome, and fitted values from this regression can be interpreted as the estimated probability of mortality. Quality for each hospital for each month is estimated using the method recommended by the Affordable Care Act, described in Chapter 3. This model fits a model of patient outcome on the patient score and a

hospital-specific random effect. A fitted estimate from this model for a hospital's patients is then used as a smoother estimate than actual patient outcomes. Risk-adjusted mortality rates then are then estimated by calculating the ratio of the fitted estimate of mortality for those patients from the random effects model to the expected mortality of those patients from the model with no random effects, and then multiplying this ratio by the actual mortality ratio in the sample.

With this estimate of quality, the main model regressed monthly AMI risk-adjusted mortality rates, *RAMR*, against monthly hospital AMI volume, a hospital-specific fixed effect, *γ*, and a time effect, δ. The time effect consists of a fixed effect for each month. This estimate shows the value of going to a larger hospital on the quality of care for the patient:

(1) $RAMR_{ht} = \alpha + \beta_1 Volume_{ht} + \gamma_h + \delta_t + \epsilon_{ht}$.

I expect to find a positive relationship between volume and quality. This is for two reasons: first, better hospitals attract more patients. Second, hospitals with more patients will get more experience and quality returns to scale. This second part of the relationship is important because it is the increase in quality due to scale and learning by doing, and is the causal effect of volume on quality. Both reasons implies a positive coefficient for β_1 on $Volume_{ht}$.

The interpretation of this estimate when hospital and monthly fixed effects are not included is very different from when these fixed effects are included. Temporal fixed effects allow the model to incorporate global shocks to volume – similar to seasonal effects. Including this as a control lessens a concern that the result is driven by seasonal codetermination of quality and of volume. Without hospital fixed effects, the effect may be dominated by between hospital relationships; in other words, that large hospitals are generally higher quality than small hospitals.

In an individual-level version of this calculation is also common in the literature and is estimated in this paper for robustness. This model estimates patient level outcomes, *M*, as a function of patient severity, monthly hospital AMI volume, a hospital-specific fixed effect, *γ*, and a time effect, δ.

(2) $M_{iht} = probit(\alpha + \beta_1 Volume_{ht} + \beta_2 E(M_{iht}) + \gamma_h + \delta_t + \epsilon_{iht})$.

Occasionally this is estimated with more than one measure of patient severity, however for simplicity this paper only uses the estimated mortality using the random forest based procedure described above.

This model doesn't match the first model because the first model does not control for expected mortality. The first model does not control for expected mortality because the dependent variable relies in part on the expected mortality rate, and thus controlling for expected mortality on the right

hand side would put that measure on both sides of the equation. However, it is possible to consider an aggregation of equation 2. An aggregate version of the dependent variable would be a hospitals total mortality rate in a month. Of the independent variables, volume and the fixed effects do not need further aggregation. An aggregate version of a patient's expected mortality is a hospital's expected mortality. This leads us to another alternate specification also here estimated in the robustness section that estimates a hospital's actual mortality rate, *MR*, as a function of its expected mortality rate, monthly hospital AMI volume, a hospital-specific fixed effect, *γ*, and a time effect, δ:

(3) $MR_{ht} = \alpha + \beta_1 Volume_{ht} + \beta_2 E(MR_{ht}) + \gamma_h + \delta_t + \epsilon_{ht}$.

Instrumental variables estimates for equations 1 and 3 can be estimated by adding another model:

(4) $Volume_{ht} = \alpha + \beta_1 z_{ht} + \beta_2 X_{ht} + \gamma_h + \delta_t + \epsilon_{ht}$

This model includes in its predictors all of the predictors of main models (none in equation 1, expected mortality rate in equation 3), indexed by the vector *X*, and fixed effects when they are included, as well as one or more instruments, *z*. These models are estimated using two stage least squares.

Equation 2 is estimated at the patient level, but the endogenous predictor is hospital-level volume. Thus, the structure of the instrumental variable model remains similar:

(5) $Volume_{ht} = \alpha + \beta_1 z_{ht} + \beta_2 E(M_{iht}) + \gamma_h + \delta_t + \epsilon_{iht}$.

Equation 2 is non-linear; a probit function is used to deal with the binary individual outcome variable, but the first stage, equation 5 is still estimated using OLS. To deal with the multi-level, non-linear nature of this system, standard errors from equations 2 and 5 are reported using a bootstrap.

In the main results, the instrument is the sum of the total cases of trauma and shock that the hospital admits in a given month. For robustness, these are considered as instruments separately. Also, for robustness, AMI volume is measured as a ratio of AMI volume in a given month to its average volume across the full 36 months.

4.4. Data

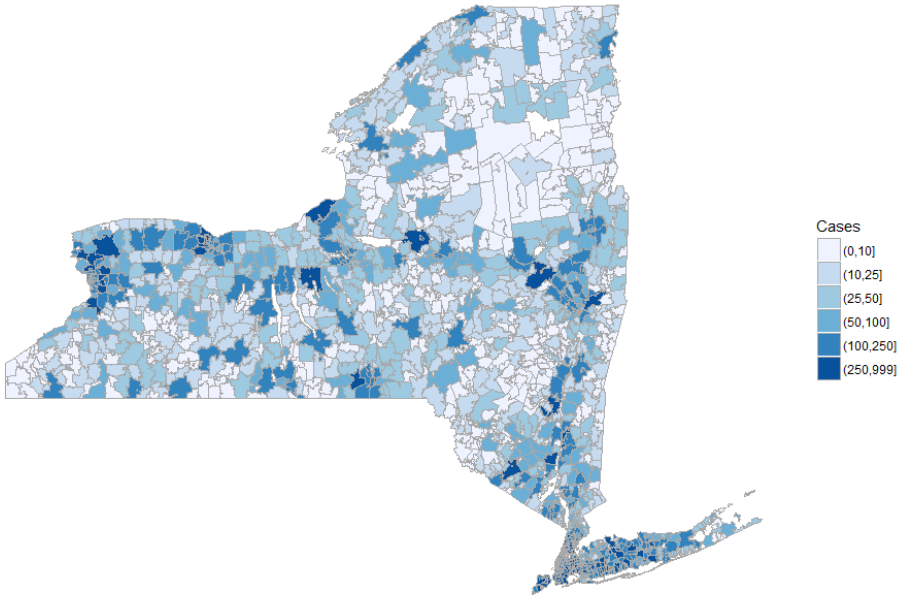
This paper uses 2005-2007 New York State Inpatient Data to create a longitudinal data set of hospital-risk adjusted mortality rates and volumes. The data was created for the Healthcare Costs and Utilization Project, a part of the Agency for Healthcare Research and Quality in the US Department of Health and Human Services. Measurement of RAMRs is restricted to AMI patients. Following the

method used by Center for Medicare and Medicaid Services (Grady et al. 2013) observations from hospitals that saw less than 25 patients in any of the three years are dropped, removing 40 facilities. Ten hospitals who served a plurality of individuals whose ZIP codes were coded as homeless and two hospitals on the border with Canada who served a plurality of individuals whose ZIP codes indicated they were foreign are dropped, resulting in a data set of 163 hospitals with 36 months of data each, and 105,842 total AMI cases[10]. Monthly volume of AMI patients, trauma patients, shock patients, and trauma and/or shock patients are recorded. Figure 4.1 is a map of New York State (with Panel 2 focusing on the part of the state in the New York City Metropolitan Area) with five-digit ZIP codes outlined and shaded by density of AMI patients. It shows that while a large proportion of the data comes from the New York City area, there is a large amount of geographic diversity.

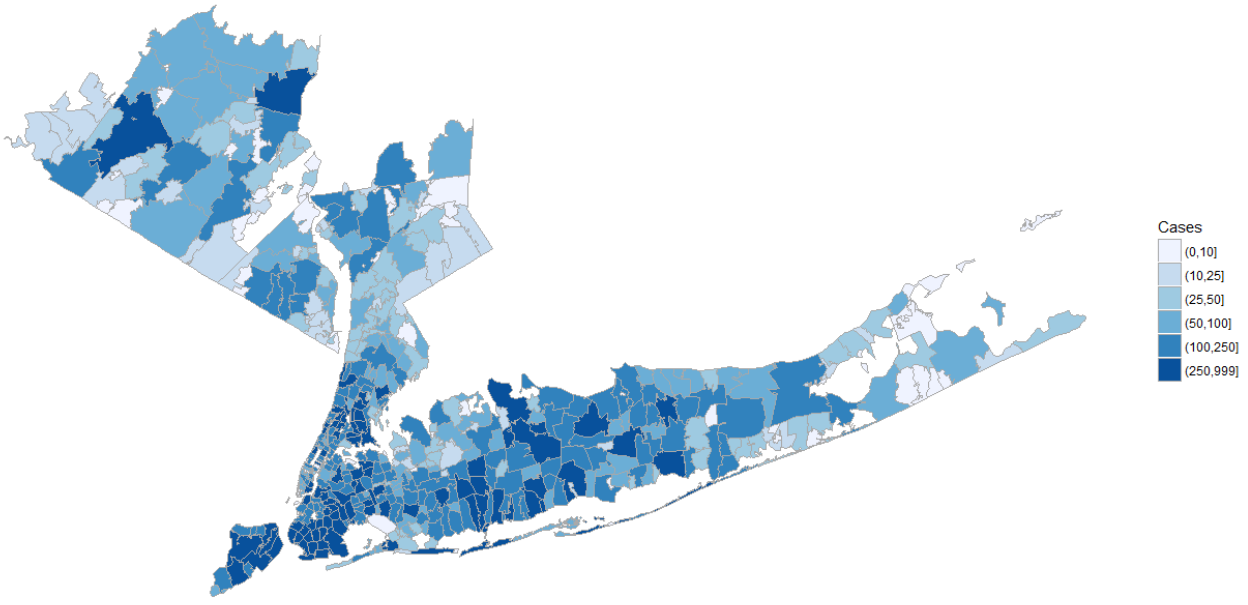
[10] Dropping hospitals who saw fewer than 25 shock or fewer than 25 trauma cases in any year was also considered. Doing this did not change the significance of the results.

Figure 4.1: Total AMI Volume of New York State and New York City MSA, Per ZIP Code, 2005-2007

Panel 1: New York State



Panel 2: New York City Metropolitan Area



Note: Data is from New York State Inpatient Database, 2005-2007.

Volume of inpatients recorded as experiencing trauma and volume of patients experiencing shock are also recorded. This data has four different common codes for shock: unspecified shock (ICD-9 code 785.50), cardiogenic shock (ICD-9 code 785.51), septic shock (ICD-9 code 785.52), and other (ICD-9 code 785.59). Volume of patients with each of these categories is considered separately and then all cases are considered together, creating 5 shock-volume variables. There are numerous types of trauma in this data, so this paper focuses on complications of shock (ICD-9 code 958), which include: air embolism as an early complication of trauma (958.0), fat embolism as an early complication of trauma (958.1), secondary and recurrent hemorrhage as an early complication of trauma (958.2), posttraumatic wound infection not elsewhere classified (958.3), traumatic shock (958.4), traumatic anuria (958.5), Volkmann's ischemic contracture (958.6), traumatic subcutaneous emphysema (958.7), other early complications of trauma (958.8), and traumatic compartment syndrome (958.9). All of these are combined, creating one shock-volume variable recording volume of patients with at least one of these categories of shock. Since these two categories are neither mutually exclusive nor universally comorbid, volume of patients experiencing trauma, shock, or both is recorded. Patients diagnosed with trauma or shock may or may not have AMI, so volume of trauma or shock cases with no AMI present are also recorded as a comorbidity. In the main analysis, hospital data are aggregated at the monthly level, considering monthly volumes and monthly risk-adjusted mortality rates. Data is also aggregated at hospital level, creating cross-sectional data consisting of monthly averages.

Quality is measured as hospital risk-adjusted mortality rate. Small, low performing hospitals are dropped, so the average mortality rate in this data, 0.0756 of patients die, is better than the average mortality rate of AMI inpatients in a more general population survey. Because higher quality hospitals are larger, the average hospital RAMR will have few large higher-performing hospitals and many small low-performing hospitals, pushing upwards the average hospital RAMR, which is 0.1274. Table 4.1 summarizes the patient admission statistics in this data, how many admissions of the three, non-mutually exclusive, key patient groups, and the average mortality rate and risk-adjusted mortality rates. There is a small decline in the mortality rate and a small increase in the risk-adjusted mortality rate of the data over the time period, but these changes are very small and not statistically significant. The number of AMI patients declined slightly, while the number of shock and the number of trauma patients rose slightly. The slopes for AMI and shock were statistically significant but only represent a change in AMI cases per hospital per month of 5.6 and in shock cases per hospital per month of 2.3.

Table 4.1: Hospital Inpatient and AMI Mortality Summary Statistics, Per Year

Statistic	N	Mean	St. Dev.	Mean for Hospitals in Top Half of AMI Volume	Mean for Hospitals in Bottom Half of AMI Volume
AMI patients	525	216.85	235.8	367	67
Shock patients	525	113.79	112.58	128	50
Trauma patients	525	11.39	25.15	34	3
Mortality rate	525	0.09	0.05	0.08	0.1
Risk-adjusted mortality rate	525	0.12	0.06	0.09	0.15

Note: Top half and bottom half of AMI Volume is based on hospitals with AMI volume greater than and less than the median value of 210. Data is from New York State Inpatient Database, 2005-2007.

Table 4.1 also reports the means for hospitals with AMI volume greater than and less than the median value of 210. Hospitals with higher AMI patient volume have lower average monthly mortality rates and RAMRs. In part this may be due to the issue of noisy estimates of RAMR and mortality rates, RAMRs and mortality rates have strict lower and upper bounds at zero and one, and that overall average RAMR and mortality rate is closer to zero. These three factors together may result in a similar character of data, that small hospitals have higher estimated RAMRs and mortality rates. To ensure this does not drive the results, models were tested with hospitals with below median volume dropped. In this case, results did not significantly change.

Some attention must also be paid to cases of AMI where the patient also is diagnosed with shock or trauma. If this were common, it could increase the correlation of AMI volume and shock and trauma volume and it could reduce the usefulness of shock and trauma volume as an instrument. However, this occurs in only 4.4% of the observations. Dropping these observations or these hospitals has little effect on the outcome. These patients do have some effect on the correlation of AMI volume and shock and trauma volume, particularly when controlling for fixed effects. Including these patients in the model, but not including them in the trauma and shock volume reduces the first stage model fit and first stage F-statistic for testing weak instruments. There is little overall change in this case in models with no fixed effects. However, in models with fixed effects, this results in a first stage F-statistic well below 10, and a very noisy estimate of the coefficient on AMI volume which is very close to zero and no longer statistically significant. While this paper drops hospitals with less than 25 AMI patients in a year, another issue of interest is hospitals with zero volume in a given month. No observations happen to have zero AMI cases, and 6.8% of observations have zero trauma or shock observations in a given month. Dropping these hospitals or hospital-months has little effect on the result.

While this paper does not focus on the determinants of hospital quality beyond volume, this paper acknowledges the role of the demographics of a hospital's patients. To approximate the demographics of a hospital's patient population, the most common home ZIP code for AMI patients at a hospital is recorded as representative of the hospital's patient population. In cases where more than one ZIP code are equally common, the ZIP code of the first patient from that group in the data is used. General demographic information about those ZIP codes are then taken from the 2007-2011 American Community Survey (data is accessed using the R package acs.r as described in Glenn 2011). The American Community Survey is an annual survey administered by the US Census Bureau as a replacement for the long form of the US Census and which first began collecting data in 2005. Data is available at the level of census tract and can be aggregated to various higher dimensions, including

ZIP code tabulation areas, which are nearly identical to ZIP codes, and which are used here. It is recommended that 5-year pooled results of the survey be used and the 2007-20011 results are used rather than the 2005-2009 because more variables are available for all New York State ZIP codes using the later series. Using this dataset, demographic information about each hospital is recorded, and that information is summarized in Table 4.2. The variables considered are the population living in the ZIP code, the median income of people living in the ZIP code, the percentage of people over the age of 25 who graduated from high school or have equivalent educational attainment, a Gini index estimate of inequality in the ZIP code, the percentage of people who work away from home who have commutes greater than 30 minutes living in the ZIP code, the percentage of the people in the ZIP code who are male, the median age of residents of the ZIP code, the percentage of the ZIP code who identify as white, and the population density of the ZIP code based on the 2010 US census. All of these variables are directly from the survey or are simple ratios of two values in the survey.

Table 4.2: Summary Demographic Information for the Plurality ZIP Code of the Hospital's Patients

Statistic	N	Mean	St. Dev.
Population	163	41,133.00	23,639.00
Median Income	163	29,187.00	10,968.00
% High School Graduates	163	0.44	0.13
Estimated Gini Index	163	0.45	0.05
% Commuters > 30 min	163	0.39	0.21
% Male	163	0.48	0.02
Median Age	163	39	5.3
% White	163	0.73	0.26
Population Density (1,000s per sq. mi)	163	6.3	11

Note: Data is from New York State Inpatient Database, 2005-2007, American Community Survey pooled 2007-2011

4.5. Results

The main goal of this part of the paper is to examine if the relationship between volume and RAMR persists when using the instrument. Table's 4.3 and 4.4 show the first and second stage of the regression described in equations 1 and 4. Table 4.3 gives the pooled estimate and the within (fixed effects) estimate. Table 4.5 gives the between estimate where all monthly observations for a hospital are averaged. The strength of the instrument is estimated using a test for weak instruments which has an f-distribution. In general, an instrument is considered weak if the F-statistic from this test is below ten. The F-statistic is above ten in all of the results in this section and in most of the results in the next section. The negative relationship between AMI volume and RAMR is consistently present in this section, giving evidence that there is a causal relationship between these two variables. To estimate the causal effect of volume on quality, this paper uses the volume of trauma and shock patients as an instrument which is correlated to the volume of AMI patients a hospital receives but does not enter a patient's or an ambulance dispatcher's choice set when a patient is sent to a hospital. Since volume of trauma or shock patients is a function of the number of trauma or shock patients in the vicinity, and is not a function of patient or ambulance choice, this instrument allows for estimation of causation in the effect of volume on quality for AMI patients, that is, how does volume improve quality. The instrumental variables estimate is likely to be smaller in magnitude than the corresponding naïve OLS and logit estimates.

Columns 1, 2, and 3 of Table 4.3 are non-instrumental variables OLS estimates using equation 1. Columns 4, 5, and 6 are instrumental variables estimates estimating equations 4 and 1 using the two-stage least squares method. In all tables, variables are scaled slightly so that coefficients are easier to read and interpret; volume is divided by 100, income is divided by $100,000, population density is divided by 1,000. The coefficient on volume in the first five columns is between -0.12 and -0.2, so if the number of patients per month increases by 10, the estimated RAMR for the hospital decreases by 1.2 to 2 percentage points. This represents a 10% to 17% decline in RAMR (average RAMR in this data is 12%). Error terms are likely to be heteroskedastic and correlated with AMI volume, so Huber-White clustered standard errors are used. This paper is most concerned with the instrumental variables estimates. The instrument gives an estimate of AMI volume which is determined by emergency medical need in the area of the hospital, and is a function of catchment area but not with quality of AMI treatment. While this exclusion restriction cannot be directly tested, the key assumption of the relationship between the instrumental variable and the instrumented variable is tested with results given in each table and more detail in Table 4.4 below.

Including both hospital and monthly fixed effects as in Column 6 results in all variation being related to random shocks to patient volume, and thus can plausibly be considered an estimate of the within hospital causal effect of volume on quality. Residual variation in volume is greatly reduced by the inclusion of hospital and monthly fixed effects which inflate the magnitude of the estimated local average treatment effect. This coefficient, -0.9, is 4.5 to 7.5 times larger than the other estimates. This relationship – a much larger instrumental variables estimate relative to OLS – is common. Three common reasons are often given, and all may exist in this example: weak instrument, measurement error, and heterogeneous treatment. As the instrument and predictor of interest are both continuous, the common measure of instrument quality compares the residual sum of squares in the instrument model (equation 4 in this case) with and without the inclusion of the instrument. In the comparison, the relevant statistic has an F-distribution, and if the F-statistic is greater than 10, the instrument is generally considered a good instrument. In Columns 4 and 5, the F-statistic for the instrument (which is the same in all three columns) is 3843 and 3771 respectively. However, in Column 6, the F-statistic is 16, still probably a good instrument, but much weaker than it was in the other columns. Since this instrument is continuous, what is called as the local average treatment effect when a binary instrument is used is sometimes interpreted as the heterogeneous treatment effect. In this case, doctors who are more likely to improve during high volume periods may be more likely to work at hospitals with higher volume. Thus, the instrumental variables estimate may be only partially valid; with a binary instrument, it might be said that the effect is locally valid. The heterogeneous treatment effect suggests that the causal relationship between volume and RAMR for hospitals which are more likely to have higher volumes. Finally, while there is very little measurement error in the data, AMI procedures may have considerable skill spillovers with other procedures, and thus AMI volume may be a noisy measure of the amount of practice physicians are getting which is applicable to the care of AMI patients. Table 8 in the robustness section presents an alternative measure of AMI volume where the instrument turns out to look a little stronger when hospital-specific and monthly fixed effects are included. In that case, the coefficient on AMI volume has a similar implication on the magnitude of the effect to the coefficients from Columns 1 through 5 of this table.

Table 4.3: Relationship between AMI Volume and AMI RAMR at Hospitals

Dependent Variable:	RAMR	RAMR	RAMR	RAMR	RAMR	RAMR
	OLS	OLS	OLS	IV	IV	IV
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)
AMI Volume (/100)	-0.153***	-0.145***	-0.123***	-0.195***	-0.178***	-0.928*
	(0.007)	(0.007)	(0.019)	(0.01)	(0.01)	(0.499)
Median Income (/$100,000)		-0.217			-0.219	
		(0.291)			(0.292)	
% High School Graduates		0.067**			0.059**	
		(0.026)			(0.027)	
Estimated Gini Index		0.106*			0.082	
		(0.057)			(0.058)	
% Commuters > 30 min		0.015			0.016	
		(0.013)			(0.013)	
% Male		-0.072			-0.074	
		(0.088)			(0.087)	
Population Density		-0.0001			0.00002	
(1,000 per sq. mi)		(0.0003)			(0.0003)	
Median Age		-0.00003			-0.00004	
		(0.0005)			(0.0005)	
% White		0.058***			0.059***	
		(0.013)			(0.013)	
Constant	0.146***	0.062	0.113***	0.154***	0.082	0.497**
	(0.003)	(0.06)	(0.013)	(0.003)	(0.061)	(0.238)
Hospital Fixed Effects	N	N	Y	N	N	Y
Monthly Fixed Effects	N	N	Y	N	N	Y
Observations	5,644	5,644	5,644	5,644	5,644	5,644
R2	0.055	0.062	0.14	0.051	0.06	0.016
F-Statistic for weak instrument test:				3843	3771	16

Note: OLS and IV estimates of the relationship between monthly AMI inpatient volume and risk-adjusted mortality rates, equation 1 and equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma and shock inpatients during the month as instruments for AMI volume. Volumes are calculated per month. Data is from New York State Inpatient Database, 2005-2007.

**p<0.1; **p<0.05; ***p<0.01*

Table 4.4 shows estimates from equation 4, and represents the first stage of the instrumental variables estimates from Table 4.3. Columns 1 and 2 without fixed effects corresponding to Columns 4 and 5 in Table 4.3, with Column 3 corresponding with Column 6 in Table 4.3 and including fixed effects. The coefficient on shock and trauma volume is more than ten times larger in Columns 1 and 2 than in Column 3. This corresponds to the instrument being weaker in that case when hospital-specific and monthly fixed effects are included. It also corresponds to the large difference in estimates of the coefficient on AMI volume in Table 4.3. In general, however, the relationships between the two kinds of volume is nearly one-to-one in Columns 1 and 2, as might be expected. In Column 3, most of the relationship is captured by the hospital fixed effect, although there remains a significant correlation between shock and trauma volume and AMI volume. The F-statistic is high in part because of the large amount of data and because the instrumental variable and the instrumented variable are both continuous. To compare this F-statistic to other F-statistics in similar studies, using the number of patients within a radius of a hospital as an instrument for CABG volume, Gowrisankaran et al. 2006 report first-stage F-test statistics of as high as 345 for some of their models.

Table 4.4: First stage of IV, Relationship between Instrument (Shock and Trauma Volume) and AMI Volume at Hospitals

Dependent Variable:	AMI Volume (/100)	AMI Volume (/100)	AMI Volume (/100)
	OLS	OLS	OLS
Independent Variables:	(1)	(2)	(3)
Shock and Trauma Volume (/100)	1.152***	1.257***	0.092**
	(0.035)	(0.041)	(0.043)
Median Income (/$100,000)		-0.224	
		(0.413)	
% High School Graduates		-0.164***	
		(0.029)	
Estimated Gini Index		-0.248***	
		(0.059)	
% Commuters > 30 min		-0.147***	
		(0.016)	
% Male		-0.363***	
		(0.085)	
Population Density		0.0004	
(1,000 per sq. mi)		(0.0004)	
Median Age		0.309***	
		(0.046)	
% White		-0.013	
		(0.013)	
Constant	0.064***	0.362***	0.453***
	(0.003)	(0.054)	(0.02)
Hospital Fixed Effects	N	N	Y
Monthly Fixed Effects	N	N	Y
Observations	5,644	5,644	5,644
R2	0.405	0.445	0.899
F-Statistic for weak instrument test:	3843	3771	16

Note: First stage of instrumental variables regressions, corresponding to equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma and shock inpatients during the month as instruments for AMI volume. Volumes are calculated per month. Data is from New York State Inpatient Database, 2005-2007.
**p<0.1; **p<0.05; ***p<0.01*

An alternative is a between estimate of the effect, where all volumes in a hospital across the entire 36 months of the data are averaged. In this case, one observation is made for each hospital, and no fixed effects are used. Table 4.5 presents this model. Column 1 is an OLS estimate of the relationship between a hospital's average monthly AMI volume and its RAMR. Column 2 uses the average monthly trauma and shock cases. In this way, this is a between hospital estimate. The coefficients in this estimate are similar to those in Table 4.3, but are about one third their magnitudes. In this case, the instrument is quite strong, giving a relatively tight bound around the coefficient on AMI volume in the second stage regression. The coefficient of 0.07 implies that comparing two hospitals, one with 10 more patients, the higher volume hospital will have 0.7 percentage points lower RAMR.

Table 4.5: Relationship between AMI Volume Ratio and AMI RAMR at Hospitals, Cross-sectional

Dependent Variable:	RAMR	RAMR
	OLS	IV
Independent Variables:	(1)	(2)
AMI Volume	-0.057***	-0.070***
(monthly average/100)	-0.007	-0.011
Constant	0.092***	0.094***
	-0.003	-0.003
Observations	163	163
R2	0.21	0.198
F-Statistic for weak instrument test:		172

Note: OLS and IV estimates of the relationship between average monthly AMI inpatient volume and risk-adjusted mortality rates, equation 1 and equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma and shock inpatients as instruments for AMI volume. Volumes are average across all the data. Data is from New York State Inpatient Database, 2005-2007.

**p<0.1; **p<0.05; ***p<0.01*

4.6. Robustness

There are a number of choices in constructing variables and in modeling strategy that may be made, and it is useful to discuss the robustness to these choices. In the case of this paper, shock and trauma volume are counted together and used as an instrument. Shock and trauma could be considered separately, which is tested in Table 4.6. For that matter, different definitions of shock and trauma are possible, which are not tested. The definition of AMI volume is not fixed, in Table 4.7 this paper looks at results if monthly volume is calculated as a ratio against average monthly volume. Tables 4.8 and 4.9 look at variations in the modelling strategy, Table 4.8 considering a patient-level model and Table 4.9 considering a different version of the hospital-level model. Some of these variations can weaken the instrument relative to the main results, but in cases where the instrument remains strong, the conclusions are generally robust to these variations.

In Table 4.6, the instrumental variables estimates from Table 4.3 are repeated but in Columns 1, 2, and 3 only the volume of shock patients are used and in Columns 4, 5, and 6 only the volume of trauma patients are used as an instrument. In Table 4.3 the total volume of patients with trauma and or shock is used as an instrument; another option is to include both volumes as two instruments in the same model. This case gives results that are nearly identical to Table 4.3 and is omitted. The first-stage regressions of the models estimated in Table 4.6 are omitted, but the F-statistic test for weak instruments is provided in the table. This result is very similar to the instrumental variables results in Table 4.3. It also shows that shock volume is more important to the result than trauma volume, but that in all cases where no hospital-specific fixed effect is used, the test for weak instruments gives a very high F-statistic. This suggests that if only trauma volume is available, it may be an effective instrument if hospital-specific fixed effects are not a part of the model. Shock may be an effective instrument in either case.

Table 4.6: Relationship between AMI Volume and AMI RAMR at Hospitals, Robustness to Different Instruments

Dependent Variable:	RAMR	RAMR	RAMR	RAMR	RAMR	RAMR
Instrument	Shock	Shock	Shock	Trauma	Trauma	Trauma
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)
AMI Volume (/100)	-0.198***	-0.179***	-1.226*	-0.170***	-0.165***	0.4
	(0.011)	(0.011)	(0.743)	(0.013)	(0.014)	(0.387)
Median Income (/$100,000)		-0.219			-0.219	
		(0.292)			(0.291)	
% High School Graduates		0.059**			0.062**	
		(0.027)			(0.026)	
Estimated Gini Index		0.08			0.091	
		(0.058)			(0.058)	
% Commuters > 30 min		0.016			0.016	
		(0.013)			(0.013)	
% Male		-0.074			-0.074	
		(0.087)			(0.088)	
Population Density		0.00003			-0.00004	
(1,000 per sq. mi)		(0.0003)			(0.0003)	
Median Age		-0.004			-0.004	
		(0.05)			(0.05)	
% White		0.059***			0.058***	
		(0.013)			(0.013)	
Constant	0.154***	0.083	0.639*	0.149***	0.075	-0.136
	(0.003)	(0.061)	(0.355)	(0.004)	(0.061)	(0.183)
Hospital Fixed Effects	N	N	Y	N	N	Y
Monthly Fixed Effects	N	N	Y	N	N	Y
Observations	5,644	5,644	5,644	5,644	5,644	5,644
R2	0.051	0.06	-0.152	0.055	0.061	0.075
F-Statistic for weak instrument test:	3672	3698	11	729	689	9

Note: OLS and IV estimates of the relationship between monthly AMI inpatient volume and risk-adjusted mortality rates, equation 1 and equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma or total volume of shock inpatients during the month as instruments for AMI volume. Volumes are calculated per month. Data is from New York State Inpatient Database, 2005-2007.

**p<0.1; **p<0.05; ***p<0.01*

Table 4.7 presents an alternative measure of AMI volume, the ratio of a hospital's AMI volume in a particular month to its average volume over all months. In this case, the instrument is very weak when hospital effects are included, but the F-statistic for the test for weak instruments is higher when hospital-specific and monthly fixed effects are included; in fact the F-statistic is higher than when raw AMI volume is used instead of the AMI-volume ratio. In this case, the coefficient on AMI volume ranges from -0.05 for the non-instrumental variables estimates to -0.14 for the instrumental variables estimate. To understand the magnitude of the IV estimate, consider a large hospital that sees 50 AMI patients in a month that sees an increase of 20%, or 10 patients. This increase in 10 patients is the same increase in volume that was considered in example in the main section in this paper. In this case, estimated RAMR in the instrumental variables estimate of Column 6 will drop 2.8 percentage points. This result is slightly larger than the estimate from Columns 4 and 5 of Table 4.3. The uncertainty of Column 6 of Table 4.3 was much higher making a comparison more difficult.

Table 4.7: Relationship between AMI Volume Ratio and AMI RAMR at Hospitals

Dependent Variable:	RAMR	RAMR	RAMR	RAMR	RAMR	RAMR
Instrument	OLS	OLS	OLS	IV	IV	IV
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)
AMI Volume Ratio	-0.054***	-0.054***	-0.056***	-7.328	-5.592	-0.135***
	(0.006)	(0.006)	(0.005)	(9.654)	(6.625)	(0.045)
Median Income (/$100,000)		-0.209			-0.209	
		(0.295)			(5.451)	
% High School Graduates		0.103***			0.103	
		(0.027)			(0.454)	
Estimated Gini Index		0.217***			0.217	
		(0.057)			(0.953)	
% Commuters > 30 min		0.01			0.01	
		(0.013)			(0.235)	
% Male		-0.064			-0.064	
		(0.088)			(1.493)	
Population Density		-0.001***			-0.001	
(1,000 per sq. mi)		(0.0003)			(0.005)	
Median Age		0.004			0.004	
		(0.05)			(0.781)	
% White		0.054***			0.054	
		(0.013)			(0.223)	
Constant	0.171***	0.026	0.127***	7.445	5.564	0.230***
	(0.007)	(0.06)	(0.011)	(9.666)	(6.69)	(0.06)
Hospital Fixed Effects	N	N	Y	N	N	Y
Monthly Fixed Effects	N	N	Y	N	N	Y
Observations	5,644	5,644	5,644	5,644	5,644	5,644
R2	0.025	0.041	0.16	-451.434	-261.606	0.112
F-Statistic for weak instrument test:				0	1	23

Note: OLS and IV estimates of the relationship between monthly AMI inpatient volume and risk-adjusted mortality rates, equation 1 and equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma and shock inpatients during the month as instruments for AMI volume. AMI Volume Ratio is equal to a hospitals AMI Volume each month divided by its average monthly volume over all months. Volumes are calculated per month. Data is from New York State Inpatient Database, 2005-2007.

**p<0.1; **p<0.05; ***p<0.01*

Table 4.8 presents estimates from a model of individual level mortality as a function of hospital monthly volume and individual patient severity as in equation 2. As with that previous table, the coefficient on AMI volume is estimated with much less certainty when hospital and monthly fixed effects are included in the model. When hospital fixed effects are not included, the coefficient, -0.0004, implies that an increase in volume by 10 patients in a month reduces the probability of an individual patient dying by 4% points. This estimate implies a slightly larger effect than in Tables 4.3 and 4.5. The IV estimate of the coefficient is slightly smaller, but nearly the same.

Table 4.8: Average Marginal Relationship between AMI Volume and Mortality for AMI Patients, Probit, Average Marginal Effect

Dependent Variable:	Patient Outcome	Patient Outcome	Patient Outcome	Patient Outcome
	GLS	GLS	IV	IV
Independent Variables:	(1)	(2)	(3)	(4)
AMI Volume	-0.000415***	-0.0000515	-0.000362***	-0.0000466
	(0.00002)	(0.003)	(0.0004)	(0.0008)
Patient Severity	0.314***	0.311***	0.315***	0.312***
	(0.003)	(0.003)	(0.003)	(0.004)
Hospital Fixed Effects	N	Y	N	Y
Monthly Fixed Effects	N	Y	N	Y
Observations	105,842	105,842	105,842	105,842
Pseudo-R2	0.32	0.33	.	.
LogLik	-18618	-18321	-501724	-392248
F-Statistic for weak instrument test:			55763	1002

Note: GLS and IV estimates of the probit relationship between monthly AMI inpatient volume and individual in-hospital mortality, equation 3 and equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma and shock inpatients during the month as instruments for AMI volume. Volumes are calculated per month. Data is from New York State Inpatient Database, 2005-2007.

**p<0.1; **p<0.05; ***p<0.01*

Table 4.9 gives estimates from equation 3. As mentioned before, this is a similar model to equation 2 estimated in Table 4.8, except that individual in-hospital mortality outcome and volume is now aggregated to the hospital-month level. These results are similar to those in Table 4.3, but the non-fixed effects estimates are about one third as large as they are in the previous table. The coefficients cannot be compared directly between Table 4.3 and Table 4.9 because AMI volume and expected mortality rates are correlated, and direct interpretation of the coefficient on AMI volume in Table 4.9 assumes that expected mortality rates are held constant. This also explains why the coefficients on AMI volume are smaller in Table 4.9 than in Table 4.3. As with other models, the F-statistic for the weak instrument test is much smaller when fixed effects are included, although it is above ten in Table 4.9 Column 6. The coefficient on AMI volume in that column is not significant, however.

Table 4.9: Relationship between AMI Volume and Mortality for AMI patients at Hospitals

Dependent Variables:	Mortality Rate	Mortality Rate	Mortality Rate	Mortality Rate	Mortality Rate	Mortality Rate
	OLS	OLS	OLS	IV	IV	IV
Independent Variables:	(1)	(2)	(3)	(4)	(5)	(6)
AMI Volume (/100)	-0.047***	-0.044***	0.022	-0.072***	-0.069***	-1.059
	(0.004)	(0.005)	(0.02)	(0.008)	(0.009)	(1.721)
E(Mortality Rate)	0.993***	0.992***	0.986***	0.987***	0.986***	0.926***
	(0.042)	(0.042)	(0.04)	(0.042)	(0.042)	(0.104)
Median Income (/$100,000)		-0.406*			-0.407*	
		(0.229)			(0.229)	
% High School Graduates		0.004			-0.002	
		(0.02)			(0.02)	
Estimated Gini Index		0.02			0.002	
		(0.044)			(0.045)	
% Commuters > 30 min		0.029***			0.030***	
		(0.011)			(0.011)	
% Male		-0.166***			-0.167***	
		(0.064)			(0.064)	
Population Density		-0.0003			-0.0002	
(1,000 per sq. mi)		(0.0002)			(0.0002)	
Median Age		0.008			0.007	
		(0.035)			(0.035)	
% White		0.024**			0.025**	
		(0.01)			(0.01)	
Constant	0.021***	0.071	-0.030***	0.026***	0.087**	0.438
	(0.004)	(0.044)	(0.01)	(0.004)	(0.044)	(0.747)
Hospital Fixed Effects	N	N	Y	N	N	Y
Monthly Fixed Effects	N	N	Y	N	N	Y
Observations	5,644	5,644	5,644	5,644	5,644	5,644
R2	0.366	0.369	0.409	0.365	0.367	0.047
F-Statistic for weak instrument test:				3945	3868	18

Note: OLS and IV estimates of the relationship between monthly AMI inpatient volume and hospital mortality rates, equation 2 and equation 4. Huber-White clustered standard errors in parentheses. IV estimates use total volume of trauma and shock inpatients during the month as instruments for AMI volume. Volumes are calculated per month. Data is from New York State Inpatient Database, 2005-2007.

**p<0.1; **p<0.05; ***p<0.01*

4.7. Conclusion

Using a novel instrument for AMI volume of shock and trauma volume, this paper gives an estimate of the effect of volume on risk-adjusted mortality rates that is of comparable magnitude to some estimates from the literature. This paper argues that this instrument is more robust to variation in geographic characteristics of a hospital's region than commonly used instruments involving the number of patients with a particular illness geographically close to the facility. The relationship estimated using ordinary least squares persists when hospital specific and monthly fixed effects are included in the model. However, the instrument becomes weaker when these effects are included, and the model no longer estimates a statistically significant relationship between instrumented volume and risk-adjusted mortality rates. However, if volume is measured as a ratio of a month's volume to the average monthly volume for that hospital in the data, the instrument is less weak, and statistical significance returns – and the magnitude is reasonable.

In many cases where administrative data is not merged with ambulance data, instruments which estimate how many cases occur within a radius of a facility are difficult to precisely estimate as patient location information may, as in this case, represent a patient's mailing home ZIP code, but this can represent a large land area and there is no verification in the data if the onset of illness occurred at home or at another place. However, data on shock and trauma volume may be generally available.

There are a number of weaknesses with using volume of trauma and shock as an instrument for AMI volume. Volume of trauma also may indicate the volume of traffic accidents and traumatic events such as occupational accidents and violent crime in an area. There may be factors which affect both the volume of traumatic events and the quality of nearby hospitals; an example being socioeconomic determinants of quality (see Nelson et al. 2002). Another factor is that hospitals with a history of high volume in general may be more likely to invest in quality. Investment in hospitals with higher general volume can lead to greater overall impact if the hospitals spend those funds on goods which are non-excludable within the hospital, that is which can improve outcomes for many illnesses. While every important socioeconomic factor cannot be perfectly accounted for, hospital-specific fixed effects are included which should limit the effect of invariant socioeconomic determinants of quality and trauma volume as well as the relationship between past volume and quality improvement. In cases where those effects are included, the instrument loses a great deal of its power. Another weakness is related to the autocorrelation of hospital volume, which could inflate instrument strength. Further work should use auto-correlation robust methods such as those suggested by Arellano and Bover (1995) and Blundell and Bond (1998) and also should consider with placebo testing, whereby the IV is applied to a setting where there would be a null causal effect.

The quality returns to volume may be due to a number of causes. Within hospital estimates could tend to represent returns to scale within the hospital and learning-by-doing. Between hospital estimates could, may be related to greater investments in hospitals during periods of high volume, even if the volume is relatively unexpected. While these estimates may be causal, if the investment-volume relationship considers measures of volume correlated to both AMI volume and to trauma and shock volume, this effect may not give an estimate that is as useful from the perspective of policy makers.

Even if volume does improve quality in a causal manner, the exact causal mechanism is not yet clear. Heusch (2009) criticizes the concept of learning-by-doing as an oversimplification. They argue that learning effects may have multiple components, including fungibility of learning across institutions and across procedures, depreciation of knowledge – possibly through forgetting, difference between learning from recent experience and from cumulative experience, and static scale effects.

The effect of volume on quality is often applied to questions of the utility of merging hospitals. This local average treatment effect may imply that the causal estimates given here of the outcomes benefit of merging facilities may be overestimated. On the other hand, the benefits estimated were significantly different from zero and persist even for low-volume facilities. Thus, while the beneficial effect of merging may be overestimated, these results suggests that merging emergency facilities which are very nearby may improve AMI outcomes if such a merger does not delay treatment for patients. Further research should seek to more precisely disentangle quality returns to scale and learning-by-doing. Both of these are important and suggest different policy implications. On the other hand, some of these returns may be captured without increasing hospital volume. If learning and experience are necessary, physicians at low volume hospitals may rotate to higher volume hospitals or use simulated cases to increase their practice and experience. Similar techniques may be developed for physicians at high volume facilities for periods when they see lower volume. Also, administrative efficiencies developed at high volume hospitals may be passed to low volume hospitals directly.

Chapter 5

5.1 Conclusion

This thesis is based on an interest in measurement and methodological issues in health economics. I have chosen to focus on questions relating to hospital quality particularly regarding risk-adjusted mortality rates (RAMRs), using a large US inpatient dataset. I have also focused on the use of a wide range of tools from econometrics and statistics that are not as common in economics. Two important contributions of this work are the exploration of the usefulness of machine learning in a health economics setting, and the use of very large inpatient datasets to address questions in health economics. The second and third chapters explore measurement and modelling issues around quality and how important those issues are in the incentivizing of high quality hospitals. The final chapter explores the role that economies of scale and learning-by-doing play in quality and introduces a novel instrument for assessing causality in the quality-volume relationship in health provision.

5.2. Summary

Improving the accuracy of estimation of hospital quality can greatly improve the efficiency of health policies aimed at improving hospital quality. The second chapter of this thesis uses a novel application of machine learning as a risk estimation of patient mortality technique in a large administrative dataset which greatly improves the ability to control for patient case mix in estimating RAMRs. It uses the technique called random forests which allows analysis of the whole data set and incorporates in its model non-linearity such as comorbidity interactions automatically. This method outperforms existing methods for mortality risk prediction. It also shows that facility rankings vary significantly between the patient risk-adjustment models presented in the chapter. Alternative models often use more limited sets of patient morbidity information by categorizing morbidities such as Elixhauser's Comorbidity Index (Elixhauser 1998), or by a variable selection process like forward stepwise logistic regression. Variation is extremely small between a risk adjustment model using only present-on-admission diagnosis codes and a model which uses all available codes. Therefore, if one's objective is to generate facility-level quality scores in data where present-on-admission information is unavailable, then the evidence suggests the preferred risk adjustment strategy is to use all available information. Given rapid increases in both the availability of data and computing power, using random forests or possibly other machine learning techniques on the fullest set of comorbidity information possible is likely to provide the best adjustment in many situations.

This chapter creates indicator variables for each International Classification of Diseases Volume 9 codes as predictors in random forests. Future research to confirm these results should also consider using disease classifications from Volume 10 of the International Classification of Diseases as well as considering its application to risk adjusted readmission rates. Common methods in the literature reduce the dimensionality of the problem by grouping morbidities into a smaller set of categories and then model mortality using generalized least squares methods based on logit or probit estimation, this work illustrates the usefulness of these new techniques in those situations. It also shows that this improvement in predictability results in significantly different estimates of hospital quality and that pay-for-performance schemes and quality reports would reflect this difference. In particular, this chapter shows that the quality of as many as one-third of hospitals would be mis-categorized in a scheme which uses RAMR to group facilities into three groups: below average, average, and above average. As such, using a machine learning method such as random forests using all POA ICD-9 codes would be the preferable method for risk adjusting patient mortality.

The third chapter continues to focus on the estimation of RAMRs as an important and commonly used measure of hospital quality. A number of different formulae are used to generate RAMRs from patient risk estimates with little concern given to the role these formula play in affecting the utility of the estimated RAMRs. The third chapter shows that RAMRs estimated using common techniques are not optimal for a number of important purposes and that non-standardized versions should be considered. Particularly, that standardizing estimates reduces their performance for four tasks health economists typically use in their work. Standardizing a hospital mortality rate is necessary when the goal is to use a hospital's actual mortality rate to estimate its RAMR, but, a hospital's predicted mortality rate is often preferred to its actual rate (Grady et al. 2013). When the hospital's mortality is predicted using hospital-specific effects, this paper argues that standardizing does not further improve the estimate. Rather, when a model with hospital-specific effects is used to predict in-hospital mortality, that prediction is itself a preferable estimate of RAMR. RAMR estimates when hospital effects are fitted with a probability model (such as in random effects models), allowing shrinkage of estimates towards the mean, further improves the estimation compared to estimating hospital effects as fixed effects.

The third chapter then presents a number of formula for comparing aggregate patient mortality risk at a hospital to estimating RAMR. It then compares the performance of the different formula in different tasks. For instance, pay-for-performance schemes are based on hospital quality ranking, and thus are most efficient when hospital quality is correctly ranked. Spearman correlation is a measure of how well the ordering of two variables, in this case RAMRs from different formula, match each

other. This work estimated the Spearman correlation of simulated true RAMR to estimated RAMR, and finds the formula with the highest correlation is not the formula commonly used by policy makers including that of the Affordable Care Act's pay-for-performance scheme, the Hospital Value Based Purchasing Plan. Depending on the scheme, this difference will play a significant role in how well a scheme aligns hospital interests to its performance. If a hospital judges that a scheme cannot discern quality signals from noise, it may reduce the willingness of a hospital to prioritize investment in quality improvement. For similar reasons, it also finds that patient choice is not always optimal when using the most popular methods of calculating RAMR.

Measurement error in RAMR when using RAMR as an outcome in regression models did result in attenuation-like bias. The chapter makes recommendations for RAMR formulae in these cases as well. When measurement is noisy, there is bias in all formula. For high levels of noise, the magnitude of the bias can be over 16% of the signal when the lowest performing formula is used, while using the preferred formula can reduce bias to less than 6% of the signal. This difference has significant implications for hypothesis testing and estimating effect sizes in modeling hospital quality.

This chapter focuses on hospital RAMR, but the results may apply to any institutional or jurisdictional groupings, such as the RAMR for a particular geographic region or facility type, or even, plausibly, for coherent groupings of patients, such as when comparing RAMR across states. The results may also apply to risk-adjusted rates of hospital readmission, which is another important measure of hospital quality.

Using shock and trauma volume as a novel instrument for AMI volume, the fourth chapter gives an estimate of the effect of volume on RAMRs that is of comparable magnitude to some existing literature, but which contradicts some of the most recent studies, such as Kim et al. (2016). This chapter argues that this instrument is more robust to variation in geographic characteristics of a hospital's region than other commonly used instruments involving the number of patients with a particular illness geographically close to the facility. The relationship estimated without an instrumental variable persists when hospital specific and monthly fixed effects are included in the model. However, the instrument becomes weaker when these effects are included, and the model no longer estimates a statistically significant relationship between instrumented volume and RAMRs. However, if volume is measured as a ratio of a month's volume to the average monthly volume for that hospital using this data, then the instrument is less weak, and the results are again statistically significant and of reasonable magnitude.

There are a number of weaknesses with using volume of trauma and shock as an instrument for AMI volume. Volume of trauma also may indicate the volume of traffic accidents and traumatic events such as occupational accidents and violent crime in an area. There may be factors which affect both the volume of traumatic events and the quality of nearby hospitals; an example being socioeconomic determinants of quality. Another factor is that hospitals with a history of high volume in general may be more likely to invest in quality. Investment in hospitals with higher general volume can lead to greater overall impact if the hospitals spend those funds on goods which are non-excludable within the hospital, that is which can improve outcomes for many illnesses. While every important socioeconomic factor cannot be perfectly accounted for, hospital-specific fixed effects are included which should limit the effect of invariant socioeconomic determinants of quality and trauma volume as well as the relationship between past volume and quality improvement. In cases where those effects are included, the instrument loses a great deal of its power.

The effect of volume on quality is often applied to questions of the utility of merging hospitals. Thus, this paper supports suggestions that merging emergency facilities which are near each other may improve AMI outcomes if such a merger does not delay treatment for patients. Further research should seek to more precisely disentangle quality returns to scale and learning-by-doing. Both of these are important and suggest different policy implications. On the other hand, some of these returns may be captured without increasing hospital volume. If learning and experience are necessary, physicians at low-volume hospitals may rotate to higher-volume hospitals or use simulated cases to increase their practice and experience. Administrative efficiencies learned at high-volume hospitals may be passed to low-volume hospitals directly.

5.3. Discussion

This thesis does leave an important question regarding the generalizability of the results. In 2013, the United States health data collection policies joined those of many other countries in adopting a new classification system for patient illness recording, shifting from volume 9 to volume 10 of the International Classification of Diseases guidelines, a fundamental variable in the second chapter. In 2014 most major provisions of the Affordable Care Act in the United States were phased in. Health care faces new and changing procedures and challenges all the time. This thesis uses data from 2005-2007, and it will be important to verify the conclusions it offers against these new conditions, and against future new conditions. Further, this thesis uses data from the largest and fourth largest US states by population, states with great regional variation that make them somewhat representative

of the country as a whole. However, next steps in research will need to focus more on this geographic variation. Rural and urban hospitals are included in the analysis, however, because the samples were largely urban, future work should assess questions of the application of this thesis to rural areas. There are many similarities, but also many fundamental differences between health care in the United States and in other countries. While these issues exist, the data is still very representative and the change in coding may not greatly affect the conclusions of the thesis. It is likely, then, that the results are generalizable to other regions and across other time periods, but it will be useful to test the robustness of its conclusions, especially to data from other countries.

These different chapters deal with questions that will be a part of the future of health economics. All three chapters take advantage of large administrative data sets. The second chapter creates a new measure of patient mortality risk which is used in the other two chapters. It also adds to the literature on the applications of machine learning to economics and health economics (Athey and Imbens 2016). The third chapter provides a unique evaluation of different methods for aggregating patient risk to hospital RAMRs, an important aspect of quality. While randomized experiments are a gold standard in understanding causal relationships, healthcare provision is an ongoing concern and many important questions cannot be addressed in a lab. The use of data analysis, including instrumental variables as is used in the fourth chapter, as well as other methods, is extremely important, especially when discussing issues about quality, where lab conditions may not give appropriate inferences. An important future goal of my work will involve using machine learning techniques in causal inference (Hill 2012, Kreif et al. 2015).

All three chapters also have important implications for health policy, particularly for incentivizing the quality of health care. Incentives for health care often come in pay-for-performance schemes, where measured hospital quality is used to modify reimbursements to hospitals for services provided. Less attention is paid to the role of cost-effectiveness in health policy incentives. Many possible policies and interventions may be considered for implementation in both public and private healthcare settings. Relative cost-effectiveness of these interventions can play a key role in how the implementation of interventions is prioritized. Better measurement of provision quality may result in more precise measures of intervention cost-effectiveness. Better measurement of cost-effectiveness increases the value of cost-effectiveness analysis as a policy tool and reduces uncertainty for policy makers comparing programs and interventions.

References

Antman, Elliott M., Mary Hand, Paul W. Armstrong, Eric R. Bates, Lee A. Green, Lakshmi K. Halasyamani, Judith S. Hochman et al. "2007 Focused Update of the ACC/AHA 2004 Guidelines for the Management of Patients With ST-Elevation Myocardial Infarction A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines: Developed in Collaboration With the Canadian Cardiovascular Society Endorsed by the American Academy of Family Physicians: 2007 Writing Group to Review New Evidence and Update the ACC/AHA 2004 Guidelines for the Management of Patients With ST-Elevation Myocardial Infarction ...." *Circulation* 117, no. 2 (2008): 296-329.

Arellano, Manuel, and Olympia Bover. "Another look at the instrumental variable estimation of error-components models." *Journal of econometrics*68, no. 1 (1995): 29-51.

Arrow, Kenneth J. "Uncertainty and the welfare economics of medical care." *The American economic review* 53, no. 5 (1963): 941-973.

Athey, Susan, and Guido Imbens. "The State of Applied Econometrics-Causality and Policy Evaluation." *arXiv preprint arXiv:1607.00699*(2016).

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M. and O'Donnell, O. Does reporting heterogeneity bias the measurement of health disparities?.*Health economics*, *17*(3), (2008) 351-375.

Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. "Assessing the accuracy of prediction algorithms for classification: an overview." *Bioinformatics* 16, no. 5 (2000): 412-424.

Bird, Sheila M., Cox David, Vern T. Farewell, Goldstein Harvey, Holt Tim, and C. Peter. "Performance indicators: good, bad, and ugly." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168, no. 1 (2005): 1-27.

Blundell, Richard, and Stephen Bond. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of econometrics* 87, no. 1 (1998): 115-143.

Breiman, Leo. Manual-Setting Up, Using, and Understanding Random Forests V4. 0. (2003) [ftp://ftpstat.berkeley.edu/pub/users/breiman]

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina. "An empirical evaluation of supervised learning in high dimensions." In *Proceedings of the 25th international conference on Machine learning*, (2008) 96-103.

Centers for Medicare & Medicaid Services (CMS), HHS. "Medicare program; hospital inpatient value-based purchasing program. Final rule." *Federal register* 76, no. 88 (2011): 26490.

Centers for Medicare & Medicaid Services (CMS), HHS. "Patient Protection and Affordable Care Act; HHS notice of benefit and payment parameters for 2015. Final rule." *Federal register* 79, no. 47 (2014): 13743.

Chu, Yu-Tseng, Yee-Yung Ng, and Shiao-Chi Wu. "Comparison of different comorbidity measures for use with administrative data in predicting short-and long-term mortality." *BMC health services research* 10, no. 1 (2010): 1.

Conway, Patrick. CMS Releases Latest Value-Based Purchasing Program Scorecard. Nov 14, 2013, The Official CMS Blog, (2013). http://blog.cms.gov/2013/11/14/cms-releases-latest-value-based-purchasing-program-scorecard/

Crandall, Marie, Douglas Sharp, Erin Unger, David Straus, Karen Brasel, Renee Hsia, and Thomas Esposito. "Trauma deserts: distance from a trauma center, transport times, and mortality from gunshot wounds in Chicago." *American journal of public health* 103, no. 6 (2013): 1103-1109.

Cutler, David M., Robert S. Huckman, and Mary Beth Landrum. The role of information in medical markets: an analysis of publicly reported outcomes in cardiac surgery. No. w10489. *National Bureau of Economic Research*, (2004).

Di Bartolomeo, Stefano, Francesca Valent, Valentina Rosolen, Gianfranco Sanson, Giuseppe Nardi, Francesco Cancellieri, and Fabio Barbone. "Are pre-hospital time and emergency department disposition time useful process indicators for trauma care in Italy?" *Injury* 38, no. 3 (2007): 305-311.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. Is more information better? The effects of'report cards' on health care providers. No. w8697. *National Bureau of Economic Research*, (2002).

Dranove, David, and Andrew Sfekas. "Start spreading the news: a structural estimate of the effects of New York hospital report cards." *Journal of health economics* 27, no. 5 (2008): 1201-1207.

Dunn, Daniel L. A comparative analysis of methods of health risk assessment. Vol. 96, no. 1. *Society of Actuaries*, (1996).

Elixhauser, Anne, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey. "Comorbidity measures for use with administrative data." *Medical care 36*, no. 1 (1998): 8-27.

Ellis, Randall P. "Risk adjustment in competitive health plan markets." *Handbook of health economics* 1 (2000): 755-845.

Epstein, Andrew J. "Effects of report cards on referral patterns to cardiac surgeons." *Journal of health economics* 29, no. 5 (2010): 718-731.

Finks JF, Osborne NH, Birkmeyer JD. “Trends in hospital volume and operative mortality for high-risk surgery.” *New England journal of medicine*. Jun 2;364(22) (2011) 2128-37.

Finlayson SR, Birkmeyer JD, Tosteson AN, Nease RF Jr. Patient preferences for location of care: implications for regionalization. *Medical care* 37 (1999) 204-209.

Fleet, Richard, and Julien Poitras. "Have we killed the golden hour of trauma?" *Annals of emergency medicine* 57, no. 1 (2011): 73-74.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, (2001).

Gaynor, Martin, Harald Seider, and William B. Vogt. "The volume-outcome effect, scale economies, and learning-by-doing." *American Economic Review* (2005): 243-247.

Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Glenn, Ezra Haber. "acs.r: an R package for neighborhood-level data from the US Census." *R: An R Package for Neighborhood-Level Data from the US Census*. Presented at the Computers in Urban Planning and Urban Management Conference, July 6, 2011.

Goldstein, Harvey, and David J. Spiegelhalter. "League tables and their limitations: statistical issues in comparisons of institutional performance." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (1996): 385-443.

Gowrisankaran, Gautam, Vivian Ho, and Robert J. Town. "Causality, learning and forgetting in surgery." *Federal Trade Commission, draft* (2006).

Grady JN, Lin Z, Wang Y, Nwosu C, Keenan M, Bhat K, Krumholz H, Bernheim S. 2013 Measures Updates and Specifications: Acute Myocardial Infarction, Heart Failure, and Pneumonia 30-Day Risk-Standardized Mortality Measure. Centers for Medicare & Medicaid Services (CMS), Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE). (March 2013).

Green DP, Vavreck L. Analysis of cluster-randomized experiments: A comparison of alternative estimation approaches. Political Analysis. 2008 Mar 20;16(2):138-52.

Grieve R, Hutton J, Green C. “Selecting methods for the prediction of future events in cost-effectiveness models: a decision-framework and example from the cardiovascular field” *Health Policy* Jun Vol 64(3) (2003) 311-324

Guru V, Tu JV, Etchells E, Anderson GM, Naylor CD, Novick RJ, Feindel CM, Rubens FD, Teoh K, Mathur A, Hamilton A, Bonneau D, Cutrara C, Austin PC, Fremes SE. “Relationship between preventability of death after coronary artery bypass graft surgery and all-cause risk-adjusted mortality rates.” *Circulation*. Jun 10;117(23): (2008) 2969-76.

Gutacker, Nils, Karen Bloor, Richard Cookson, Chris P. Gale, Alan Maynard, Domenico Pagano, José Pomar, and Enrique Bernal-Delgado. "Hospital Surgical Volumes and Mortality after Coronary Artery Bypass Grafting: Using International Comparisons to Determine a Safe Threshold." *Health Services Research* (2016).

Hannan EL, Wu C, Ryan TJ, Bennett E, Culliford AT, Gold JP, Hartman A, Isom OW, Jones RH, McNeil B, Rose EA, Subramanian VA. “Do hospitals and surgeons with higher coronary artery bypass graft surgery volumes still have lower risk-adjusted mortality rates?” *Circulation*. Aug 19;108(7) (2003) 795-801.

Hofer TP, Hayward RA. “Identifying poor-quality hospitals. Can hospital mortality rates detect quality problems for medical diagnoses?” *Medical Care*. Aug;34(8) (1996) 737-53.

Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality, Rockville, MD. (2014), http://hcupnet.ahrq.gov/.

Hill JL. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics. 2012.

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. "Random survival forests." *The annals of applied statistics*(2008): 841-860.

Joynt, Karen E., and Ashish K. Jha. "Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program." *Jama* 309, no. 4 (2013): 342-343.

Karnon, Jonathan, Orla Caffrey, Clarabelle Pham, Richard Grieve, David Ben-Tovim, Paul Hakendorf, and Maria Crotty. "Applying Risk Adjusted Cost-Effectiveness (Rac-E) Analysis To Hospitals: Estimating The Costs And Consequences Of Variation In Clinical Practice." *Health economics* 22, no. 6 (2013): 631-642.

Kim, Woohyeon, Stephen Wolff, and Vivian Ho. "Measuring the Volume-Outcome Relation for Complex Hospital Surgery." *Applied health economics and health policy* (2016): 1-12.

Koetsier A, de Keizer NF, de Jonge E, Cook DA, Peek N. "Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: a simulation study." *Critical Care Medicine*. Jun;40(6) (2012) 1799-807.

Kreif, Noémi, Richard Grieve, Iván Díaz, and David Harrison. "Evaluation of the Effect of a Continuous Treatment: A Machine Learning Approach with an Application to Treatment for Traumatic Brain Injury." *Health economics* 24, no. 9 (2015): 1213-1228.

Krumholz, Harlan M., Ralph G. Brindis, John E. Brush, David J. Cohen, Andrew J. Epstein, Karen Furie, George Howard 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 and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation." *Circulation* 113, no. 3 (2006): 456-462.

Krumholz, Harlan M., Yun Wang, Jennifer A. Mattera, Yongfei Wang, Lein Fang Han, Melvin J. Ingber, Sheila Roman, and Sharon-Lise T. Normand. "An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction." *Circulation* 113, no. 13 (2006): 1683-1692.
Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. "Improving propensity score weighting using machine learning." *Statistics in medicine* 29, no. 3 (2010): 337-346.

Lee, Kris CL, Kannan Sethuraman, and Jongsay Yong. "On the Hospital Volume and Outcome Relationship: Does Specialization Matter More Than Volume?" *Health services research* 50, no. 6 (2015): 2019-2036.

Li, Pengxiang, Michelle M. Kim, and Jalpa A. Doshi. "Comparison of the performance of the CMS Hierarchical Condition Category (CMS-HCC) risk adjuster with the Charlson and Elixhauser comorbidity measures in predicting mortality." *BMC health services research* 10, no. 1 (2010): 1.

Mark BA, Harless DW, McCue M. "The impact of HMO penetration on the relationship between nurse staffing and quality." *Health Economics*. Jul;14(7) (2005) 737-53.

McCoy, C. Eric, Michael Menchine, Sehra Sampson, Craig Anderson, and Christopher Kahn. "Emergency medical services out-of-hospital scene and transport times and their association with mortality in trauma patients presenting to an urban Level I trauma center." *Annals of emergency medicine* 61, no. 2 (2013): 167-174.

McKay, Niccie L., and Mary E. Deily. "Cost inefficiency and hospital health outcomes." *Health economics* 17, no. 7 (2008): 833-848.

Nelson, Alan R., Brian D. Smedley, and Adrienne Y. Stith, eds. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press, 2002.

Newgard, Craig D., Eric N. Meier, Eileen M. Bulger, Jason Buick, Kellie Sheehan, Steve Lin, Joseph P. Minei, Roxy A. Barnes-Mackey, Karen Brasel, and ROC Investigators. "Revisiting the "Golden Hour": An evaluation of out-of-hospital time in shock and traumatic brain injury." *Annals of emergency medicine* 66, no. 1 (2015): 30-41.

Newgard, Craig D., Robert H. Schmicker, Jerris R. Hedges, John P. Trickett, Daniel P. Davis, Eileen M. Bulger, Tom P. Aufderheide et al. "Emergency medical services intervals and survival in trauma: assessment of the "golden hour" in a North American prospective cohort." *Annals of emergency medicine* 55, no. 3 (2010): 235-246.

Park, H. K., H. S. Ahn, S. J. Yoon, H. Y. Lee, J. M. Hong, S. W. Lee, and H. J. Hann. "Comparing risk-adjusted hospital mortality for CABG and AMI patients." *Journal of international medical research* 33, no. 4 (2005): 425-433.

Pine, Michael, Harmon S. Jordan, Anne Elixhauser, Donald E. Fry, David C. Hoaglin, Barbara Jones, Roger Meimban, David Warner, and Junius Gonzales. "Enhancement of claims data to improve risk adjustment of hospital mortality." *Jama* 297, no. 1 (2007): 71-76.

Pitches DW, Mohammed MA, Lilford RJ. "What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature." *BMC Health Services Research*. 7:91. (2007).

Pons, Peter T., Jason S. Haukoos, Whitney Bludworth, Thomas Cribley, Kathryn A. Pons, and Vincent J. Markovchick. "Paramedic response time: does it affect patient survival?" *Academic Emergency Medicine* 12, no. 7 (2005): 594-600.

Racz MJ, Sedransk J. "Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes." *Journal of the American statistical association*. 105(489) (2010) 48–58

Ramanarayanan, Subbu. "Does practice make perfect: An empirical analysis of learning-by-doing in cardiac surgery." *Available at SSRN 1129350* (2008).

Rosenthal, Meredith B., and Richard G. Frank. "What is the empirical basis for paying for quality in health care?" *Medical Care Research and Review* 63, no. 2 (2006): 135-157.

Ross, Joseph S., Sharon-Lise T. Normand, Yun Wang, Dennis T. Ko, Jersey Chen, Elizabeth E. Drye, Patricia S. Keenan et al. "Hospital volume and 30-day mortality for three common medical conditions." *New England Journal of Medicine* 362, no. 12 (2010): 1110-1118.

Schull, Michael J., Marian J. Vermeulen, and Therese A. Stukel. "The risk of missed diagnosis of acute myocardial infarction associated with emergency department volume." *Annals of emergency medicine* 48, no. 6 (2006): 647-655.

Shahian, David M., Gregg S. Meyer, Elizabeth Mort, Susan Atamian, Xiu Liu, Andrew S. Karson, Lawrence D. Ramunno, and Hui Zheng. "Association of National Hospital Quality Measure adherence with long-term mortality and readmissions." *BMJ quality & safety* 21, no. 4 (2012): 325-336.

Silber, Jeffrey H., Paul R. Rosenbaum, Tanguy J. Brachet, Richard N. Ross, Laura J. Bressler, Orit Even-Shoshan, Scott A. Lorch, and Kevin G. Volpp. "The Hospital Compare mortality model and the volume–outcome relationship." *Health services research* 45, no. 5p1 (2010): 1148-1167.

Smith, Peter C., and Andrew D. Street. "On The Uses of Routine Patient-Reported Health Outcome Data." *Health Economics* 22, no. 2 (2013): 119-131.

Southern, Danielle A., Hude Quan, and William A. Ghali. "Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data." *Medical care* 42, no. 4 (2004): 355-360.

Street, Andrew. "How much confidence should we place in efficiency estimates?." *Health economics*12, no. 11 (2003): 895-907.

Swaroop, Mamta, David C. Straus, Ogo Agubuzu, Thomas J. Esposito, Carol R. Schermer, and Marie L. Crandall. "Pre-hospital transport times and survival for hypotensive patients with penetrating thoracic trauma." *Journal of emergencies, trauma, and shock* 6, no. 1 (2013): 16.

Wang, Justin, Jason Hockenberry, Shin-Yi Chou, and Muzhe Yang. "Do bad report cards have consequences? Impacts of publicly reported provider quality information on the CABG market in Pennsylvania." *Journal of health economics* 30, no. 2 (2011): 392-407. Harvard

Wilde, Elizabeth Ty. "Do emergency medical system response times matter for health outcomes?" *Health economics* 22, no. 7 (2013): 790-806.