Predicting Hospital Safety: A Comparison of Four Models

May 5, 2010

J. Scott Ashwood
H. John Heinz III College
Carnegie Mellon University

Committee Members:
Martin Gaynor, H. John Heinz III College (chair)
Amelia Haviland, RAND Corporation and H. John Heinz III College
Daniel Nagin, H. John Heinz III College

Abstract

Choosing the best hospitals or avoiding the worst hospitals is important to patients, physicians, and insurers. There is a wealth of publicly reported quality data to use in making those choices, but that data is presented in ways that may conceal important information. We develop three models of how individuals might use the available data in choosing a hospital and present a model that relies on a group-based alternative to the current reporting methods. We compare these four models on their ability to both predict hospital performance on quality and to identify the best and worst hospitals using three measures of patient safety calculated for California hospitals 1997-2006. We find that there are no clear winners in predicting the levels of quality, but we do find evidence that group-based models may be better at identifying the worst hospitals.

1. Introduction

There are many sources of information on the levels of hospital patient safety in the United States. All of them report safety for individual hospitals in the form of rankings or for populations of hospitals in the form of average trends. The two main reasons for public reporting of patient safety levels are to provide consumers with information they can use in deciding where to seek care and to provide policy makers and researchers with information they can use for safety improvement (Marshall, et al., 2000).

Prior research demonstrated that a group-based models approach can identify groups of hospitals following distinct trajectories of safety rates over time that an average trend or annual rankings will miss (Ashwood, 2009). One of the conclusions of that research was

that policy makers and researchers may be able to identify poor performing hospitals more easily through a group-based approach than through traditional approaches. This in turn could lead to overall safety improvement by focusing efforts on the hospitals that need the most help. We focus in this paper on how individuals might use the publicly reported patient safety information in choosing hospitals.

Patients who are able to chose which hospital they go to for care, physicians who are choosing where to practice as well as where to send their patients, and public and private insurers who are choosing which hospitals to contract with may be very interested in how a hospital will perform in the future. They are interested in going to or contracting with the safer hospitals and avoiding the dangerous ones. Publicly reported quality information is one source they can use to make that choice. While there is research into whether patients and physicians use publicly available quality information (Faber, et al., 2009; Marshall, et al., 2000; Werner and Asch, 2005), that literature does not evaluate how they might use that information to predict hospital performance. We provide some alternative prediction models individuals could use to predict hospital performance.

In this paper, we will introduce four models that can be used to predict hospital performance on quality and will test each of them on three measures of patient safety. Three of the models we chose are those that are implied by the ways in which patient safety is currently reported. They reflect the theoretical decision models that patients, physicians, and insurers could use when they choose a hospital based on publicly available performance reports. In addition, we include group-based models of trajectories to see whether they add insight in the context of prediction. We will use a training dataset consisting of seven years of patient safety levels for each hospital to fit our theoretical models. We will then evaluate the performance of each of the models in predicting future patient safety rates over three additional years.

We find that none of the models stands out as the best in all situations; however, the group-based models may do better at identifying the worst hospitals and one of the hospital level models may do better at identifying the best hospitals. For the data that we

use, each approach performs well at predicting actual safety levels for some measures and not as well for others. In the end, it may not be the level of performance but the safety category that matters. Knowing whether a hospital is safe or dangerous may matter more than knowing whether its safety rate is 2% better than another's. The evidence and conclusions we present may only apply to the data we use for our analyses and may not hold up outside our sample.

2. Background

In 2000, the Institute of Medicine (IOM) released a report on patient safety that included an estimate that adverse events occur in 2.9 to 3.7 percent of hospitalizations (Institute of Medicine, 2000). The report estimates that every year, 50,000 to 100,000 people die from preventable adverse events that occur during hospital stays in the United States, and that the estimated cost of these events is $17 billion (Institute of Medicine, 2000). The IOM report stimulated a great deal of research into the causes of adverse events with the goal of improving patient safety. Measuring and reporting patient safety have become important components of healthcare policy and research (Greenberg et al. 2009, Provonost et al. 2006, Zhan et al. 2005).

There are several databases tracking patient safety outcomes in the U.S. that are maintained by public and private organizations (Faber et al 2009; Greenberg et al. 2009; Zhan et al. 2005). Government agencies and private organizations report patient safety levels and trends in the U.S. (CDC 2004; Joint Commission 2009; AHRQ 2007a). For example, the Joint Commission reports an increase in the number of reported wrong-side surgeries from under 5 in 1997 to more than 70 in 2006 (Joint Commission 2009), and the CDC reported the distribution of blood stream infections 2006-2007 by hospital ward (Edwards et al. 2008).

In addition to reporting levels of patient safety, public and private organizations create safety grades for individual hospitals and report rankings on patient safety measures (The Leapfrog Group 2008; Health Grades Inc. 2009; Hospital Compare 2009). These grades

and rankings are based on hospitals' patient safety records in that year or based on the average record over three years. For example, Health Grades Inc. provides a list of hospitals by state that are top performers in 2008 on patient safety based on their average performance on a selected set of AHRQ PSI's over the past three years (Health Grades Inc. 2009).

Public reporting was originally designed to improve safety through competition for safety seeking patients and through publicly shaming poor performing hospitals. In theory better informed consumers will choose higher quality hospitals, and hospitals that perform poorly will improve their safety levels in order to climb in the rankings. There is some evidence to support the notion that patients consider quality in their choice of hospitals. The economic literature on hospital demand includes two studies that provide evidence that patients consider both distance and quality in deciding among hospitals (Howard, 2005; Tay, 2003). The health services literature provides mixed evidence on how patients and physicians use publicly available quality information is mixed (Faber, et al., 2009; Marshall, et al., 2000; Werner and Asch, 2005). Lisa Schwartz and her colleagues found that Medicare beneficiaries thought that publicly available quality information was important in their choice on where to go for surgery but relied primarily on the recommendation the referring physician in making their decision (Schwartz, et al., 2005). They suggest that quality information should be targeted at physicians who can then direct their patients towards higher quality surgeons and hospitals. However, Schneider and Epstein found that a surgeon quality report card had minimal influence on referrals by physicians in Pennsylvania (Schneider and Epstein, 1996). The main reason why was a belief by physicians that the specific measures in the report card were not credible, though physicians in their study felt that quality measures are important in general. Publicly available measures of healthcare provider quality are considered important by both physicians and patients, but the way in which that information is presented or the specific measures used have led to lower actual response to quality reports (Faber, et al., 2009; Marshall, et al., 2000).

If patients and physicians use publicly reported quality information, then the current methods for reporting that information may lead them to the wrong conclusions. Prior research provided evidence that the current methods for reporting safety levels may be missing some important aspects of the distribution of patient safety (Ashwood, 2009). That research provided evidence that there are multiple groups of hospitals following distinct trajectories of safety rates 1997-2006. Population average trends and cross-sectional ranking miss this important information. That means analysts or policy makers who are studying and working to improve patient safety may be missing key information. It also means that patients and physicians may not be able to accurately identify the best or worst providers.

For this paper, we will take it as given that patients and physicians use publicly available quality information in their decision on which hospitals to use. In order to determine whether the group-based models approach can help improve patient and physician decision-making, we set up four models of how quality information can be used to predict future hospital safety. Three of the models are implied by the current methods for reporting safety, and one is the group-based model of trajectories. We then test each model's ability to predict future performance on a set of patient safety outcome measures that we have computed for hospitals in California.

3. Data

The primary data source for this paper is the annual hospital discharge data for the state of California, 1997-2006. This public dataset is maintained by the California Office of Statewide Health Planning and Development (OSHPD). It includes a record for each hospital discharge during a single calendar year from a California licensed hospital. Each discharge record contains information about the reason for the hospital admission as well as what happened during the stay. The information included on the discharge record is mainly collected for billing purposes and is not as rich as the patient's medical record. That means there may be valuable, medically relevant information missing from the discharge record. Despite this lack of all of the medically relevant information, billing

records, like the discharge records analyzed here, are easier to acquire and analyze, and they are used widely in healthcare research (Romano et al. 2003; Zhan and Miller 2003a). The patient safety measures I use are designed to be used on discharge data.

The California discharge files include discharges for four types of facilities: general acute care, acute psychiatric, chemical dependency recovery, and psychiatric health. This paper focuses on general acute care hospitals. These are the most common type of hospital in the state and the one most people are familiar with. There are an average of 488 facilities of all types in each year in the data, with an average of 353 general acute care hospitals. The general acute care hospitals account for 91% of the discharges per year on average.

4. Performance measure

4.1 AHRQ Patient Safety Indicators

We use the AHRQ Patient Safety Indicators (PSI's). The PSI's cover several aspects of hospital care and have been widely used in research and policy. They have been used to study the impact of hospital structure and safety culture on outcomes (Rivard et al, 2009; Singer et al, 2009), the impact of adverse events on the cost of hospitalizations (Zhan and Miller, 2003b), and are also included as a performance in pay for performance proposals (Lindenaeur et al, 2007).

The PSI's are designed to "screen for problems that patients experience as a result of exposure to the healthcare system and that are likely amenable to prevention by changes at the system or provider level." (AHRQ, 2007b, p.2) They are designed to be used with hospital discharge abstract data, the kind of data that all states routinely collect from hospitals. There is not as much information in the discharge record as there is in the actual medical chart, for example, but chart review for 3 million discharges per year would be prohibitively expensive. As a result of the limitations of administrative data, the quality indicators analyzed here do not provide a definitive measure of patient safety.

They can only be used "to provide *indicators* of health care quality that can serve as the starting point for further investigation." (AHRQ, 2007b, p.12).

4.2 Computing Raw Rates

There are 20 provider level measures (Table 1) that cover a variety of adverse events from the relatively minor (selected infections due to medical care) to fatal (death in low-mortality DRG's). AHRQ has created a set of programs for the statistical package SAS that can generate hospital level rates for each of the 20 PSI's. This software for computing the measures has gone through several revisions as AHRQ has made changes to the methods for identifying adverse events as well as to the risk adjustment process. One of the most important revisions to the process involves adding consideration of whether a condition is present on admission. The present on admission flag can improve estimates of patient safety events as well as improve discharge level risk adjustment (Bahl et al. 2008; Glance et al., 2006a; Glance et al., 2006b). We use version 3.1 of the AHRQ software to generate hospital level PSI rates which includes the use of present on admission flags.

Table 1. AHRQ Patient Safety Indicators

1 - Complications of Anesthesia
2 - Death in Low-Mortality DRGs*
3 - Decubitus Ulcer
4 - Failure to Rescue
5 - Foreign Body Left During Procedure*
6 - Iatrogenic Pneumothorax
7 - Selected Infections Due to Medical Care
8 - Postoperative Hip Fracture
9 - Postoperative Hemorrhage or Hematoma
10 - Postoperative Physiologic and Metabolic Derangements*
11 - Postoperative Respiratory Failure*
12 - Postoperative Pulmonary Embolism or Deep Vein Thrombosis
13 - Postoperative Sepsis*
14 - Postoperative Wound Dehiscence
15 - Accidental Puncture or Laceration
16 - Transfusion Reaction*
17 - Birth Trauma – Injury to Neonate
18 - Obstetric Trauma – Vaginal with Instrument
19 - Obstetric Trauma – Vaginal without Instrument

20 - Obstetric Trauma – Cesarean Delivery*

* Unable to compute rates using CA discharge data

Each discharge is examined for eligibility for each of the indicators (added to the denominator) as well as whether it appears one of the events actually occurred (added to the numerator). Each PSI has its own set of eligibility and event requirements. For example, PSI 3 (Decubitus Ulcer) identifies all cases that have a diagnosis code indicating decubitus ulcer. Cases that have a length of stay less than 5 days or are admitted from a long term care facility are excluded since decubitus ulcers generally take longer than 5 days to develop and it is possible that an ulcer developed at the long term care facility rather than the hospital.

Three of the PSI's require identifying whether a hospitalization involved an elective surgery. The California discharge data does not have the information necessary to identify whether surgery is elective, so we are not able to compute rates for Postoperative physiologic and metabolic derangement (PSI 10), Postoperative respiratory failure (PSI 11), and Postoperative sepsis (PSI 13). In addition to these three PSI's, we are unable to compute trajectories for four PSI's that are extremely rare. They are Death in low mortality DRG's (PSI 2), Foreign body left during procedure (PSI 5), Transfusion reaction (PSI 16), and OB Trauma – C-section (PSI 20). The seven indicators that we cannot compute are identified with an asterisk in Table 1 above.

4.3 Adjustment Process

Once raw PSI rates are computed for each hospital, they can be risk-adjusted to account for the risk profile of each hospital's patients. Some hospitals see sicker patients on average than others. Using the demographic characteristics of the hospital's patients as well as the mix of diagnosis categories (DRG's), the PSI software generates an expected rate for each hospital. This rate is generated by applying coefficients estimated from a regression of patient demographics and disease profiles on PSI rates for a national sample of hospitals. In addition to the coefficients, the PSI software includes average national

rates for each PSI. The hospital's predicted rate is compared to this national average. If the hospital's predicted rate is higher than the national average, then the hospital has a more severe case mix than average, and their observed rate is adjusted downward. If their predicted rate is lower, then they have a less severe case mix than average, and their observed rate is adjusted upward. This new rate is the risk-adjusted rate and represents "the rate the provider would have if it had the same case-mix as the reference population given the provider's actual performance." (AHRQ, 2007, p.65) Finally, the risk-adjusted rates are smoothed in order to remove excess variation for smaller hospitals. These final, smoothed rates are used for the analyses reported in this paper.

4.4 Selection of Three PSI's

We focused on three PSI's for our analysis. To select the three from among the 13 possible PSI's, we looked for PSI's that are not extremely rare and that are recommended or endorsed by any national hospital quality organizations or federal agencies. We also looked for PSI's with consistent definitions over time and with high signal ratios as reported by AHRQ. The signal ratio is the proportion of the total variation the rate that is due to systematic differences among hospitals (AHRQ, 2007). It is computed by analyzing the variance of residuals from a logistic regression predicting the presence of a PSI on each discharge and identifying the proportion of the variance that is due to the hospital (signal). This is called multivariate signal extraction because the logistic model includes several risk adjustment covariates, including age, gender, DRG, and comorbidities (AHRQ, 2002). Measures with higher signal ratios are considered more reliable because they are not as sensitive to randomness. The higher correlation in residuals within a hospital suggests that hospital differences account for the differences in rates rather than chance randomness (Hofer et al., 1999). Table 2 contains the information We used to make our final selection.

Table 2. Information used to select three PSI's for analysis

PSI	Avg Rate 1997-2006	AHRQ signal ratio	Recommended/Endorsed/Used By		
			HQA	NQF	CMS Premiere Hospitals
1	0.04	75.7			
3	2.67	85.5			
4	252.32	66.6		Yes	
6	0.43	79.9		Yes	
7	1.29	70.8			
8	0.04	67.1			
9	1.99	8.6			Yes
12*	2.87	72.6		Yes	
14	1.96	35.6	Yes	Yes	
15*	2.26	82.9	Yes	Yes	
17*	4.33	97		Yes	
18	145.38	69.9			
19	33.32	86.4			

* Indicates final selection

We chose PSI 12 (Postoperative Pulmonary Embolism or Deep Vein Thrombosis), PSI 15 (Accidental Puncture or Laceration), and PSI 17 (Birth Trauma). PSI's 1, 6, and 8 are very rare with average annual rates/1000 discharges of 0.04, 0.43, and 0.04 respectively. PSI's 9 and 14 have low signal ratios, according to AHRQ. PSI 1 has been dropped from the current version of the PSI software, and PSI 4 has a major change in definition in the version immediately after the one we use. Of the remaining PSI,s 12, 15, and 17 are all endorsed by either the Hospital Quality Alliance or the National Quality Forum, or both. Both of these are respected national hospital quality organizations.

5. Models

We chose models for comparison that reflect some theoretical models of how patients might predict the performance of hospitals using publicly reported patient safety levels. Three of the models are based on the ways patient safety levels are currently described,

and one of the models is the group-based model of trajectories. There are three broad classes of models that we consider: hospital level, population level, and group level.

The hospital level models reflect a method for predicting hospital performance that is implied by the hospital ranking systems currently used. The implication is that a hospital's future performance can be best modeled by looking only at that hospital's own performance in the past. There are two hospital level models: average (the average performance over the past 3 years is the best predictor of future performance), and trend (hospital's own trend over time is the best predictor).

The population level model reflects a method for predicting performance that is implied by the average trend reports of safety levels. Hospitals are expected to follow the population level rate of change over time, though they may have their own starting points. We use a standard fixed effects model that includes a single estimated trend over time and hospital indicators that allow the intercept to vary by hospital.

Finally, we use group-based models of trajectories to estimate group level trends as well as the form of each group's trajectory.

We will estimate each model using rates from 1997-2003, generate predicted values for the 2004-2006 time period and compare the performance of each model using both the root mean squared prediction error and the percentage of the time that each model accurately predicts hospitals' relative performance levels.

5.1 Model 1: Hospital's average performance in last three years

The first hospital model that we consider is a model that takes the average PSI rate from the last three years as a prediction of future performance. There are several hospital rankings that base rankings on performance in a single year (Hospital Compare 2009) or on average performance over some period of time (The Leapfrog Group 2008; Health Grades Inc. 2009). If consumers of hospital services look at these rankings as an

indication of hospital performance, then they may assume that hospital performance is captured by the rankings. If so, then they may assume that a hospital will perform at roughly the same level as implied by the ranking.

We tested the validity of using the hospital ranking approach as a method for predicting hospital performance by using the average PSI rate for the hospital in 2001-2003 as a prediction for the rates in 2004-2006. The equation for performance in year t is:

$$\hat{Y}_{it} = \frac{\sum_{j=2001}^{2003} Y_{ij}}{3} \qquad (1)$$

Where t is 2004, 2005, or 2006.

5.2 Model 2: Hospital trend over seven years

Model 1 does not allow for the possibility of change over time for hospitals that are improving or getting worse. To address this possibility, we estimate a hospital level trend for each hospital for 1997-2003 and then project that trend into the 2004-2006 time period as an estimate of hospital performance. The equation for performance in year t is:

$$\hat{Y}_{it} = \hat{\alpha}_i + j\hat{\beta}_i \qquad (2)$$

Where t is 2004, 2005, or 2006, $\hat{\alpha}_i$ is the intercept, j is some vector of years 1997-2003 that may include higher order values, $\hat{\alpha}_i$ and $\hat{\beta}_i$, are estimated using data from 1997-2003.

5.3 Model 3: Population average trend with hospital fixed effects

Another common way to report safety levels for hospitals is the population average trend. The implication for performance is that there is a single average trend that all hospitals follow. We have modeled this as a fixed effects panel model, where we estimate a single

average trend but allow individual hospitals to have different starting points. The equation for performance in year t is:

$$\hat{Y}_{it} = \hat{\alpha}_i + j\hat{\beta} \qquad (3)$$

Where $\hat{\alpha}_i$ is a hospital specific intercept and $\hat{\beta}$ is the vector of population average trend coefficients. $\hat{\alpha}_i$ and $\hat{\beta}$ are estimated using data from 1997-2003. This model falls into the class of models used to handle unobserved effects in panel models (Wooldridge, 2002; Cameron and Trivedi, 2005), though we are motivating this model differently. We are specifying model 3 in this way to reflect the type of information available to patients about the safety levels in each hospital rather than to deal with unobserved hospital characteristics that could bias our estimates of $\hat{\beta}$.

5.4 Model 4: Group-based models of trajectories

The group-based trajectories modeling technique (Nagin 2005) takes advantage of longitudinal encounter data for individual hospitals, to determine whether groups of hospitals appear to follow distinct trends in safety outcomes. Trajectory modeling supports both an estimate for the number of distinct groups (or clusters) of hospitals implied by the data, as well as an estimate for the actual trajectory in outcomes for each group. Traditional growth models can be used to model outcomes over time, but they assume either that there is a single average trend, or that each hospital is following its own distinct trend, like models 2 and 3 above. Group-based models allow for the possibility that there are multiple trends, but provide a way to combine several individual trends together into meaningful groups (Nagin 2005; Muthén and Muthén 2000). As a result, groups of hospitals with stable rates of outcomes can be distinguished from those with either increasing or decreasing rates.

The group-based trajectory methodology is based on maximum likelihood estimation (MLE): A likelihood function is specified and the values of the parameters are chosen to maximize the likelihood function. The estimated parameters include both the size of each

group as well as the form of the group trajectories (e.g. constant, linear, quadratic, etc.). In this case, the likelihood of an individual hospital's observed patient safety rates over time is modeled as a function of average group rates over time. The dependence of an individual hospital's PSI rates from year to year is captured by the group average trends. Conditional on group membership, individual variation around the group average is assumed to be independent. For example, a group of hospitals may be estimated to follow a decreasing trend over time. That captures the dependence in rates for a single hospital that is a part of the group. The rate in year three should be lower than the rate in year two. The fact that the hospital may have a rate that is higher than the group average rate in one year and lower than the group average rate in the next is assumed to be independent. The specific form of this individual variation around the group average can be modeled several ways. The three most common ways are to model the variation as normal, Poisson, and binary logit (Nagin, 2005). The likelihood function itself depends on which of these forms is chosen.

The number of groups and the shape of the trajectories are inputs for the likelihood function. As a result, the appropriate number of groups and the appropriate form of the trajectory (constant, linear, quadratic, or cubic) must be carefully chosen. To estimate the number of groups, Nagin recommends using a combination of statistical score (Bayesian Information Criterion score) and subject matter knowledge (Nagin, 2005).

The Bayesian Information Criterion (BIC) score, derived from the maximized likelihood score, increases as the maximized log likelihood increases, but imposes a penalty for increasing the number of parameters (Raftery, 1995). Note that a separate set of parameters for the size of the group and the shape of the trajectory is included for each group. The BIC score will eventually decrease as the number of groups increases.

Selecting the number of groups that corresponds to the highest BIC score is one way to choose the number of groups. However, adding more groups to improve the BIC score may not improve the understanding of the patterns of patient safety in the data. It is important that the choice is meaningful in the context of the problem (Nagin, 2005). For

example, a trajectory model with two groups of hospitals might identify a large group that has a low, constant rate of adverse safety outcomes, and a second group that has a high and increasing rate. A trajectory model with three groups might have a higher BIC score than does the two-group model. Nevertheless, if the third group is simply identified by splitting the low, constant group into two pieces, then little insight has been gained, especially if the new third group is estimated to have very few members. The important conclusion is captured by the two group model: namely, the fact that most hospitals have a low and stable rate of adverse outcomes, while others have an increasing rate of outcomes on the same measure. In sum, relying solely on the BIC score to estimate the appropriate number of groups in trajectory modeling may lead to choosing too many groups to be useful.

Once the number of groups is estimated, the pattern of PSI rates over time must be estimated. That is, the degree of the polynomial in time must be estimated. The form is important because it describes whether and how rates change over time. We used a SAS procedure, PROC TRAJ, to estimate the group-based trajectories models (Jones et al., 2001).

As in the models above, we used PSI rates for 1997-2003 to estimate the individual trajectories and group sizes. In addition, PROC TRAJ provides the posterior probability of group membership for each group for each hospital. To compute the predicted PSI rates 2004-2006, we use the weighted average of the k trajectories, where the weights equal the posterior probabilities. The equation for performance in year t is:

$$\hat{Y}_{it} = \sum_k (\hat{\pi}_{ik} * j\hat{\beta}_k) \qquad (4)$$

Where $\hat{\pi}_{ik}$ is the posterior probability that hospital i belongs to group k and $j\hat{\beta}_k$ is the average PSI trend 1997-2003 for group k.

6. Analysis

We fit each of the four models described above to the data we have for 1997-2003. We then use the formulas (1-4) above to predict hospital performance in each year 2004-2006. We compare the ability of each model to predict hospital performance in two ways. We compute the root mean squared prediction error for each model and look for the lowest value. We also rank hospitals on observed and predicted PSI rates and compute the percentage of agreement between the predicted and observed data for selected quantiles of the distribution. We look for the highest percent agreement.

6.1 PSI rates over time

Each of the three PSI's has a different pattern in rates over time. The average rate for PSI 12 (Figure 1) is increasing 1997-2006, PSI 15 (Figure 2) is bouncing around a lot, and PSI 17 (Figure 3) is decreasing on average. In each of the three figures below, the red line indicates the end of period used to fit the data.

Figure 1: Average rate for PSI 12 (Postoperative PE or DVT)

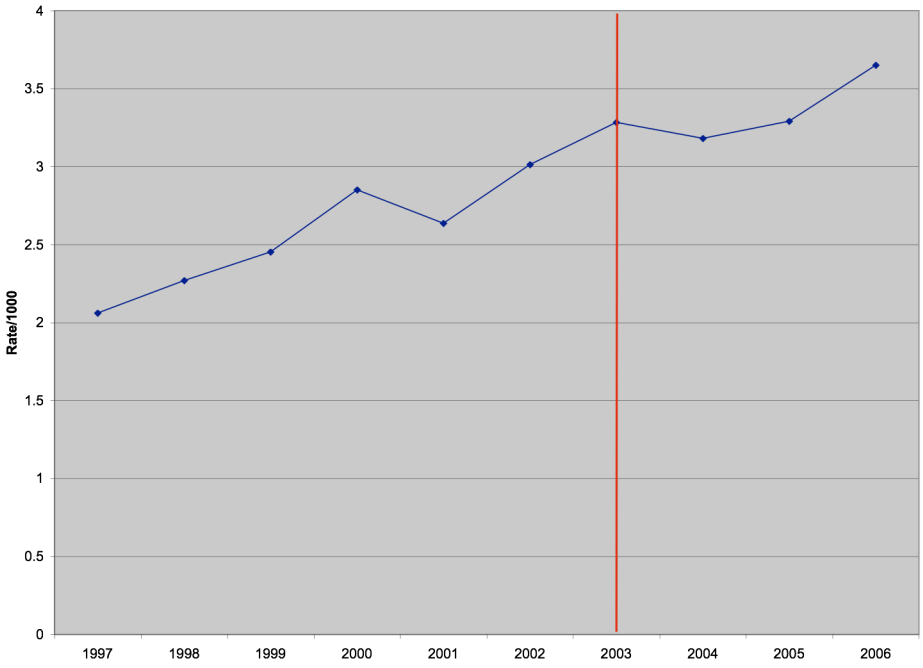


Figure 2: Average rate for PSI 15 (Accidental puncture or laceration)

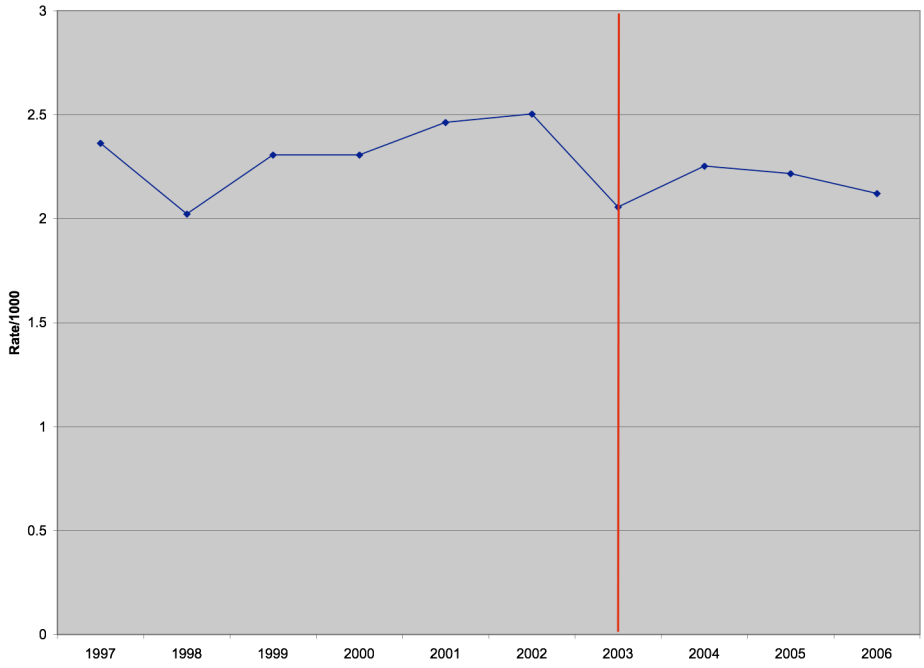
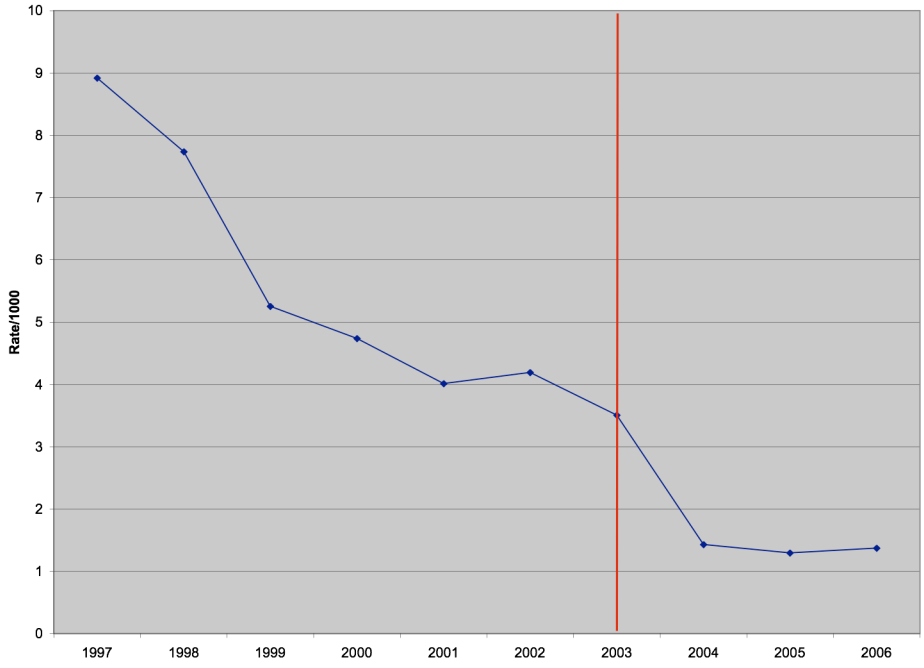


Figure 3: Average rate for PSI 17 (Birth Trauma)



In addition to differences in the average trends, the three PSI's differ in their group trends. Figures 4-6 below show the estimated trajectories for each PSI 1997-2003. PSI 12 is estimated to have five groups, PSI is estimated to have four groups, and PSI 17 is estimated to have two groups. All three PSI's have large groups following relatively constant trajectories and smaller groups following more dynamic trajectories. In each figure below, the lines represent the predicted average trends for the associated groups, and the percentages in parentheses represent the estimated group sizes.

Figure 4: 5 Trajectories for PSI 12 (Postoperative PE or DVT)

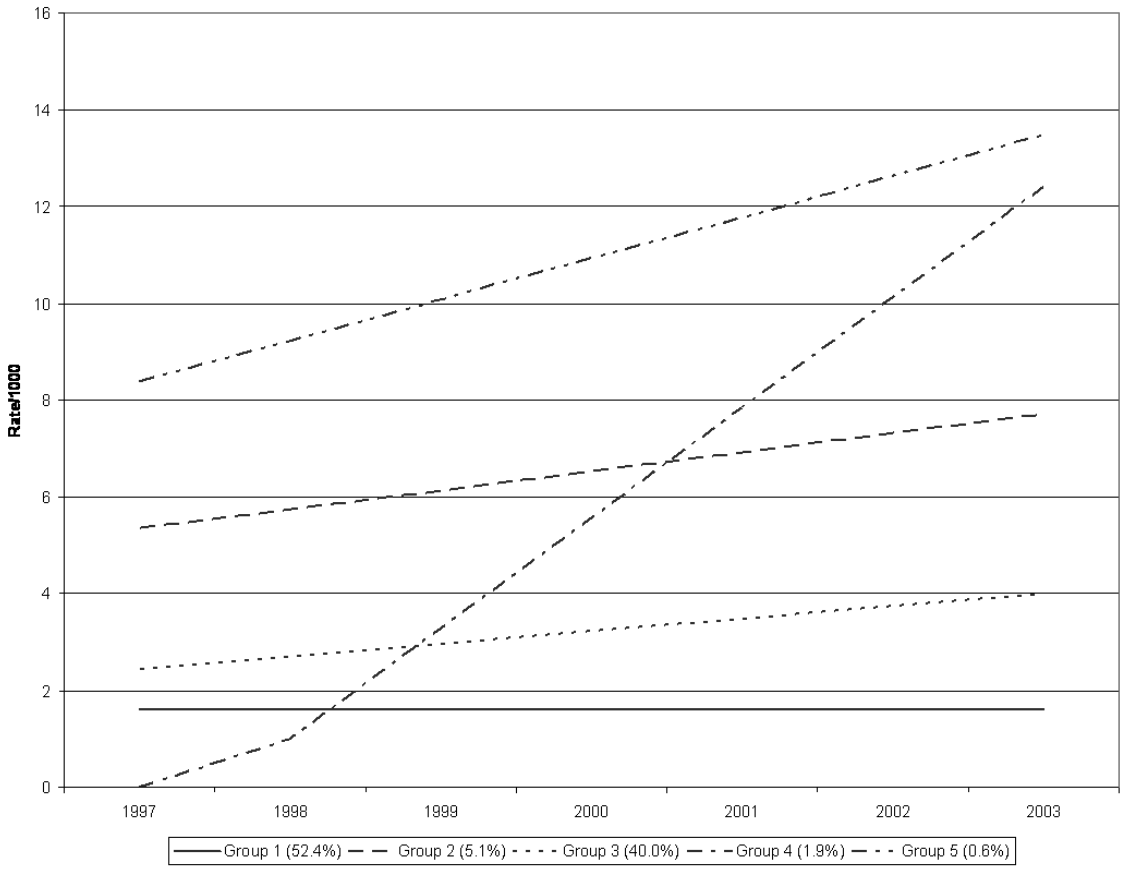


Figure 4 shows that there is evidence for five groups for PSI 12. Three of those groups are following trajectories with positive slopes, like the overall average trend in Figure 1,

but these groups are small (estimated to include 7.6% of the hospitals). Most of the hospitals appear to be following one of two constant trajectories.

Figure 5: 4 Trajectories for PSI 15 (Accidental puncture or laceration)

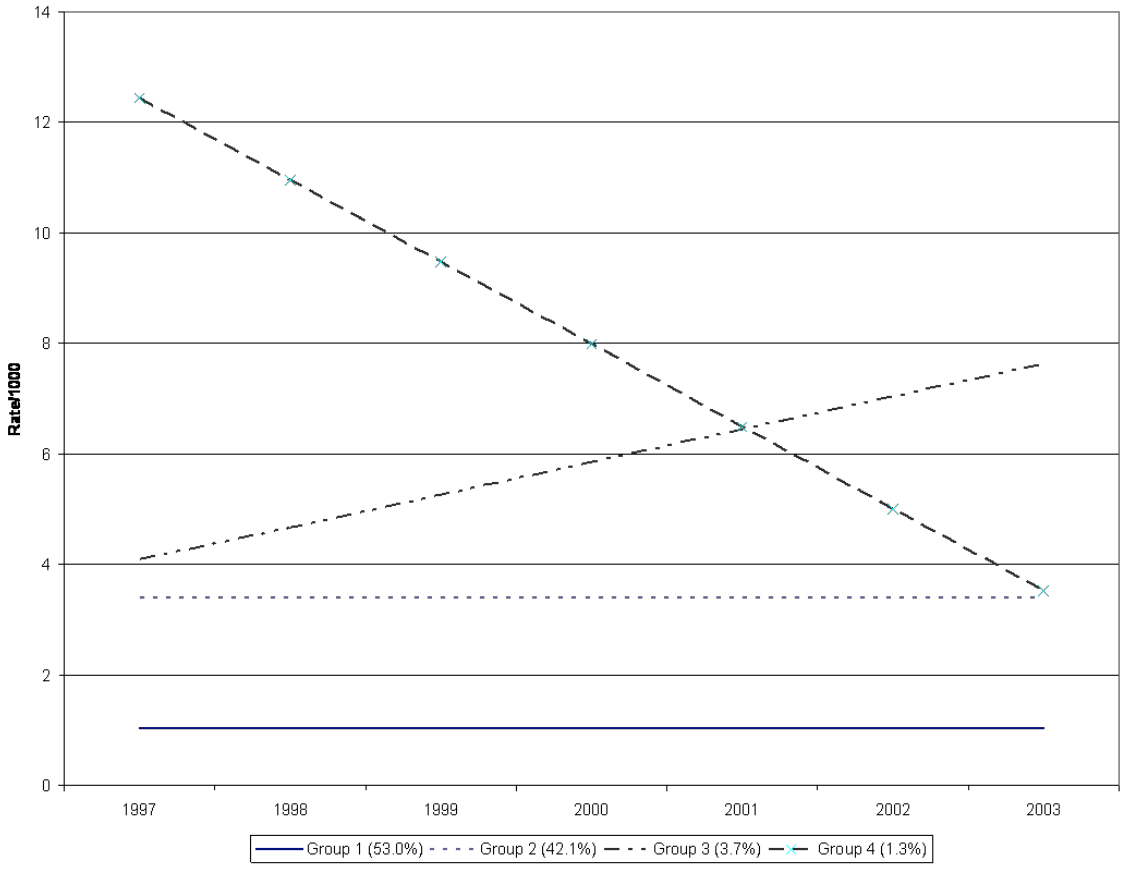


Figure 5 provides evidence of four groups for PSI 15. One group (estimated to include 3.7% of the hospitals) appears to have an increasing rate, and one group (1.3%) appears to have a decreasing rate. Like PSI 12, there appear to be two large groups with constant rates.

Figure 6: 2 Trajectories for PSI 17 (Birth Trauma)

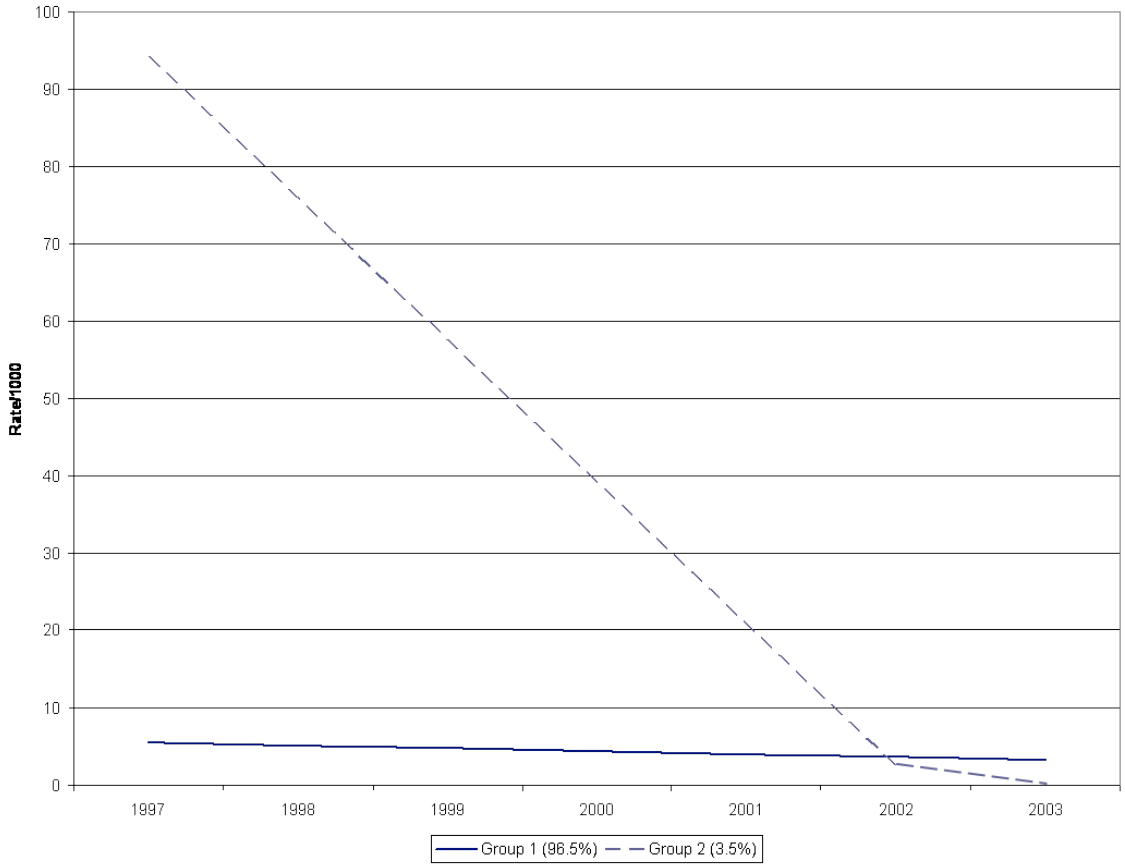


Figure 6 provides evidence that there are two groups for PSI 17. One small group is experiencing a decline in birth trauma rates, but most hospitals appear to have a low and constant rate. The downward trend in the population average (Figure 3) appears to be driven mainly by a small group of hospitals.

6.2 Model selection and prediction

In each of the linear models and the group-based trajectories model, we tried different functional forms to fit rates over time. In each case we looked at both the fit to the data for 1997-2003 (training data) and the fit to the data for 2004-2006 (testing data). Our measure of model fit was the root mean squared error (RMSE) for the training data and the root mean squared prediction error (RMSPE) for the testing data. All of our models, with the exception of the three-year average, are models that minimize the squared-error

loss. Thus, it is appropriate to use the RMSE and the RMSPE as methods for evaluating the fit of each model.

Normally, anyone trying to predict future performance only has the observed data and would only be able to consider the fit of a model to the observed data. Since we have divided our data into two groups, we are able to compare the fit in both sets. That allows us to choose the model specification that does the best in our testing data.

In general we found that models that fit the training data best do not do as well in the testing data. Simpler models outperformed more complicated models at prediction. This is an example of the tradeoff between over-fitting and over-smoothing (Wasserman, 2004). In the case of PSI 12 for example, the linear specification for the population average model captures the overall average increase 1997-2003 in Figure 1. A higher order model would capture the rise and dip in the average rate in 2000 and 2001, and would fit the training data better. In order to fit variation in rates in the middle well, the model might cause the predicted rates to increase at a very high rate in the validation data. Figure 1 shows that the rate does increase in the 2004-2006 period, but only after dropping slightly from the 2003 rate. A cubic specification that fits the training data well will project the rate to be farther and farther away from the observed gradual decline. The linear specification more closely models the observed data 2004-2006 with a slight increase. The fact that the line estimated in the training data is less subject to the year to year variation proves important in this case.

We have overcome the problem of over-fitting our data by looking to the testing data. In essence, we are looking into the future to find the best model. In a more realistic setting, without the ability to look into the future, cross-validation is the recommended way to decrease the likelihood of over-fitting the observed data.

For each model, we generated predicted values for each year 2004-2006 using the estimated coefficients from the models run on the 1997-2003 data. In the cases in which the predicted values were less than 0, we recoded them to equal 0. This recoding had the

largest impact on the hospital specific and group-based trajectories models. It affected none of the records for PSI 12, 1.2% of the records for PSI 15, and 3.9% for PSI 17.

After generating predicted values for each year, we average them for the three year period and compare them to the average observed rate for that period. The conclusions are the same whether we use the average or look at each year individually.

6.3 Root mean squared prediction error

In order to compare models on their ability to predict safety levels, we use the root mean squared prediction error (RMSPE). We looked at this measure in two ways, for the entire population, and by the groups identified by the group-based approach.

Table 3: Root mean squared prediction error overall

	Hospital Level Models			
	3 year avg	Trend	Population Trend	Group Trend
PSI 12	2.391	3.207	2.175	2.708
PSI 15	1.718	2.107	1.720	1.719
PSI 17	5.443	6.460	7.194	1.628

Table 3 above contains the RMSPE values for each of the models. The lowest RMSPE's are highlighted for each PSI. A different model does the best by this measure for each of the PSI's. The population trend model (model 3) does best for PSI 12, the three year average (model 1) does best for PSI 15, and the group trend model (model 4) does best for PSI 17.

We also looked at the RMSPE by the groups identified by the group-based trajectories models. Figures 4-6 above provide evidence that there are different trends within the data, and it appears that the group-based trajectories approach does well in predicting for some of these trends. Table 4 contains RMSPE values for each model by group. To get these numbers, we assigned each hospital to its most likely group, based on the posterior

probability of group membership. We then took the square root of the average squared prediction error within each group.

Table 4: Root mean squared prediction error by group

PSI 12

		Hospital Level Models			
Group (n)	**Classification**	**3 year avg**	**Trend**	**Population Trend**	**Group Trend**
1 (166)	Stable/Low	2.118	2.409	1.943	1.915
2 (15)	Getting Worse	3.543	3.463	3.552	3.258
3 (121)	Stable	2.132	2.711	2.059	1.982
4 (5)	Getting Worse	7.259	14.965	4.782	13.927
5 (2)	Getting Worse	3.180	3.652	3.115	4.347

PSI 15

		Hospital Level Models			
Group (n)	**Classification**	**3 year avg**	**Trend**	**Population Trend**	**Group Trend**
1 (167)	Stable/Low	1.417	1.560	1.388	1.381
2 (136)	Stable	1.535	1.897	1.620	1.627
3 (11)	Getting Worse	4.519	6.622	3.462	4.734
4 (5)	Getting Better	3.849	2.984	5.731	1.904

PSI 17

		Hospital Level Models			
Group (n)	**Classification**	**3 year avg**	**Trend**	**Population Trend**	**Group Trend**
1 (223)	Stable/Low	5.321	6.572	3.231	1.636
2 (8)	Getting Better	8.141	0.955	34.686	1.427

In addition to presenting the RMSPE values, Table 4 includes a classification for the groups. The classification is based on examining Figures 4-6 above. The stable/low groups are groups that appear to be doing well for the 1997-2006 time period. The groups that are getting worse may be hospitals that may need special attention. It appears that the group-based trajectories model does better than all of the other models at

predicting PSI rates for the stable/low groups. In addition, it does better for some of the smaller, more dynamic groups, though the population average model does better at predicting the future performance of the groups that are getting worse.

For example, in PSI 12, the group-based models perform better on both of the stable groups and one of the groups that is getting worse. It does worse on the other two groups that are getting worse. The population trend model does better for these two groups. It may be that the increasing population trend in Figure 1 is driven by these two smaller groups. They may both continue to experience increasing rates 2004-2006, but may move back towards the population average rate. The group-based models would project these hospitals to increase at too high a rate. Note how poorly the group-based model performs for group 4 which has the highest slope.

Despite doing worse than other models when all of the hospitals are pooled, the group-based models appear to do better for most of the hospitals when the data is analyzed by group. Most of the hospitals in our data appear to follow the stable and stable/low trajectories, and the group-based models do better than the others at predicting for these groups. That may be because the group-based models avoid the noisiness of the individual models and the sensitivity to outliers of the population based model. Most of the hospitals are in the stable/low or stable groups for which the group-based model does well. For predicting the actual levels of performance for the hospitals that are getting worse, the group-based models do not do as well as the population average models.

It is important to note that tables 3 and 4 are providing point estimates for the RMSPE's. Since we are analyzing almost all of the general acute care hospitals in California 1997-2006, the point estimates are valid for statements about that population. If we want to generalize beyond this population, then we need some estimate of how far apart these measures would be in a larger population. For example, the RMSPE for the group-based models for the PSI 12 stable/low group is lower than any of the others, but 1.915 may not be different than the 1.943 RMSPE for the group-based model. When we compute confidence intervals for the RSMPE's in tables 3 and 4 we find that all of these models

perform at the same level for all of the PSI's and groups. That is, all of the confidence intervals overlap with each other and several span the entire range of point estimates.

6.3 Predicting relative performance

In addition to evaluating the ability of these models to predict actual performance levels, we also evaluated how these models do at predicting relative performance. The hospital ranks generally group hospitals into top performers, and CMS's pay for performance demonstration provided financial rewards for the top percentiles of hospitals and penalized the lowest percentiles (CMS, 2006). Patients and physicians may ultimately be most interested in whether a hospital is safe or dangerous.

We evaluated each model on its ability to predict the top 10% and bottom 10% of the hospitals 2004-2006. After generating the predicted rates, we average them over the three years and order the hospitals from highest to lowest average predicted rate. We then identify the hospitals in the top 10% (rate <= 10^{th} percentile) and the bottom 10% (rate >= 90^{th} percentile). Since our measures are accident rates, we consider the "top" hospitals to be those with the lowest rates and the "bottom" to be those with the highest rates. We also rank hospitals by their average observed rates 2004-2006 and identify the top and bottom based on percentiles from the observed distribution.

Once we have classified hospitals as top or bottom based on both the predicted and observed rates, we compute the percentage of the time the predicted and observed classifications agree. Table 5 contains the percent agreement for each of the models.

The hospital only models do not appear to do well for predicting the levels of performance, but the three-year average model does appear to do better than the others at identifying the best hospitals. The three-year average approach (model 1) does best at predicting the best hospitals for both PSI 12 and 15. The hospital trend approach (model 2) does best for predicting the worst hospitals for PSI 17. Despite doing well at predicting the levels of performance certain types of hospitals, the population average

models does not do well at predicting relative performance. The group-based trajectories approach (model 4) does best at predicting the best hospitals for PSI 17 and does best at predicting the worst hospitals for PSI's 12 and 15.

Table 5: Predicting "Best" and "Worst" hospitals (percent agreement)

Predicting the "Best" Hospitals

	Hospital Level Models			
	3 Year Average	Trend	Population Trend	Group Trend
PSI 12	89.64	86.41	87.06	89.00
PSI 15	89.34	83.70	88.71	86.83
PSI 17	86.15	83.55	85.28	87.01

Predicting the "Worst" Hospitals

	Hospital Level Models			
	3 Year Average	Trend	Population Trend	Group Trend
PSI 12	89.64	90.29	90.94	91.59
PSI 15	88.71	88.09	88.71	89.97
PSI 17	86.15	87.01	83.55	86.15

The group-based models identify groups of hospitals that appear to be getting worse over time and that as a result may not be considered safe, but they do not predict the actual safety levels of the hospitals in these groups very well. However, despite the fact that the group-based approach does worse than the population average model at predicting the average PSI rates for the groups that are getting worse, it may still be better at identifying hospitals that will be the worst performers 2004-2006. The group-based approach may predict rates that are too high for these groups and therefore perform badly when looking at the RMSPE. Nonetheless, when the relative performance is considered, the group-based approach does well because it has identified the hospitals that will have the highest rates 2004-2006 even if it has greatly overestimated those rates.

Once again, we must treat these conclusions with caution. While the point estimates presented in table 5 are different, confidence intervals on these estimates suggest that we cannot generalize beyond our data. As in the case of the RMSPE estimates, the confidence intervals on the percent agreement values overlap one another and span the entire range of estimated values.

7. Conclusions

If patients and physicians use publicly available data in the process of choosing a hospital, then they must use a prediction model that relies on hospital information only or population average information only, or some combination of the two. We developed three models that could be used by patients and physicians and compared them on their ability to predict the levels of performance and ability to identify the best and worst hospitals. The hospital level models do not do well at predicting the levels of hospital safety, but the three-year average models do well at predicting the best hospitals. The population average models do perform well at predicting the levels of performance for certain types of hospitals but do not do well at predicting which hospitals will be the best or the worst.

We have presented evidence that group-based models of trajectories can provide some insight that the more common ways of reporting patient safety miss. They do not do well at predicting the level of performance on average but do perform better than the other models for subsets of hospitals. In particular, they do better at predicting levels of performance for hospitals that have low accident rates for the entire time period. In addition, the groups-based models appear to do better at identifying the hospitals that will perform the worst. Furthermore, the group-based models provide evidence that while there may technically be a group of hospitals that can be identified as the best, that group is probably indistinguishable from a very large group of hospitals that are all performing well. Thus, the ability to identify the smaller set of hospitals that is getting worse may be more useful to consumers who are choosing among hospitals.

Patients and physicians are not the only groups who may benefit from the insights gained from a group-based approach. Insurance providers, both public and private, may be interesting in predicting hospital performance over time. In addition to the interest in contracting with better hospitals, they profile hospitals they have already contracted with and are exploring ways to link payments to performance. Most of the models they use rely on some method for profiling providers. Profiling methods that focus on physicians have grouped physicians together by specialty and patient mix. The group level rates are used to adjust individual physician scores on quality. Grouping hospitals for comparability is more complicated because most hospitals perform a wide range of services to a wide range of patients. The group-based approach presented here could be useful, especially if that grouping increases the ability to predict performance. The evidence presented above suggests that is possible.

Patients, physicians, and insurers are currently unable to use group-level information in making their decision, but it appears that such information could be useful. It may be especially useful for identifying the less safe hospitals. The evidence we provide in this paper is only suggestive for hospitals outside our data; however, we have presented evidence that among the hospitals we analyze there are large groups of safe hospitals and smaller groups of hospitals that are not performing well on safety. Both of those conclusions are impossible to reach using the current methods for reporting safety.

8. Acknowledgments

I would like to thank my committee for their excellent advice and guidance throughout the process of producing this paper.

9. References

Agency for Healthcare Quality and Research (AHRQ). 2002. *Measures of Patient Safety Based on Hospital Administrative Data – The Patient Safety Indicators.*

Rockville, MD, U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality. [accessed on April, 13, 2010]. Available at http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=hstechrev&part=A17071.

Agency for Healthcare Quality and Research (AHRQ). 2007a. *2007 National Healthcare Quality Report.* Rockville, MD, U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality, 136pp.

Agency for Healthcare Quality and Research (AHRQ). 2007b. *Guide to Patient Safety Indicators. Version 3.1.* Rockville, MD, U.S. Department of Health and Human Services, Agency for Healthcare Research and Quality. [accessed on Oct, 25, 2008]. Available at http://www.qualityindicators.ahrq.gov/downloads.htm.

Ashwood, J.S. 2009. "Applying Group-based Trajectories Methodology to Measures of Patient Safety" Heinz first paper. Draft available at http://www.heinz.cmu.edu/faculty-and-research/faculty-profiles/faculty-details/index.aspx?faculty_id=297

Bahl, V., M.A. Thompson, T.Y. Kau, H.M. Hu, and D.A. Campbell, Jr.. 2008. "Do the AHRQ Patient Safety Indicators Flag Conditions That Are Present at the Time of Hospital Admission?" *Medical Care* 46 (5): 516-22.

Cameron, A. Colin and Pravkin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York, NY: Cambridge University Press.

Centers for Disease Control (CDC). 2004. "National Nosocomial Infections Surveillance (NNIS) System Report, data summary from January 1992 through June 2004, issued October 2004." Atlanta, Georgia, Division of Healthcare Quality Promotion, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Public Health Service, US Department of Health and

Human Services. [accessed on April 21, 2009]. Available at http://www.cdc.gov/ncidod/dhqp/nnis_pubs.html

Edwards JR, Peterson KD, Andrus MA, Tolson JS, Goulding JS, Dudeck MA, et al. 2008. "National Healthcare Safety Network (NHSN) report, data summary for 2006 through 2007, issued November 2008." *Am J Infect Control* 36: 609–26.

Faber M., M. Bosch, H. Wollersheim, S. Leatherman, and R. Grol. 2009. "Public Reporting in Healthcare: How Do Consumers Use Quality-of-Care Information?" *Medical Care* 47 (1): 1-8.

Glance L.G., A.W. Dick, T.M. Osler, and D.B. Mukamel. 2006a. "Does Date Stamping ICD-9-CM Codes Increase the Value of Clinical Information in Administrative Data?." Health Services Research 41 (1): 231-51.

Glance L.G., A.W. Dick, T.M. Osler, and D.B. Mukamel. 2006b. "Accuracy of Hospital report Cards Bassd on Administrative Data." *Health Services Research* 41 (4): 1413-37.

Greenberg M.D., A.M. Haviland, H. Yu, and D.O. Farley. 2009. "Safety Outcomes in the United States: Trends and Challenges in Measurement." *Health Services Research* 44 (2): 739-55.

Health Grades, Inc. 2009. *The Sixth Annual HealthGrades Patient Safety in American Hospitals Study*. Available at http://www.healthgrades.com/media/dms/pdf/PatientSafetyInAmericanHospitalsStudy2009.pdf

Howard, D.H. 2005. "Quality and consumer choice in healthcare: Evidence from kidney transplantation." *Topics in Economic Analysis and Policy* 5 (1) Article 24.

Hospital Compare. 2009. U.S Department of Health and Human Services (HHS). Hospital Compare website. Available at http://www.hospitalcompare.hhs.gov

Institute of Medicine. 2000. *To Err is Human: Building a Safer Health System.* Edited by L.T. Kohn, J.M. Corrigan, and M.S. Donaldson. Washington, DC: National Academy Press.

Joint Commission. 2009. "Sentinel Events Trends Reported by Year." [accessed on April 19, 2009]. Available at http://www.jointcommission.org/SentinelEvents/Statistics/

Jones, B.L., D. Nagin, and K. Roeder. 2001. "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories." *Sociological Research and Methods* 29: 374-393.

The Leapfrog Group. 2008. "2008 Leapfrog Top Hospitals." [accessed on April 19, 2009]. Available at http://www.leapfroggroup.org/media/file/2008_Top_Hospital_Release.pdf

Lindenaeur, P.K et al. 2007. "Public Reporting and Pay for Performance in Hospital Quality Improvement." *New England Journal of Medicine* 356: 486-496.

Martin, M.N., P.G. Shekelle, S. Leatherman, and R.H. Brook. 2000. "The Public Release of Performance Data: What Do We Expect to Gain? A Review of Evidence." *Journal of the American Medical Association* 283 (14): 1866-1874.

Muthén B. and L.K. Muthén. 2000. "Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling With Latent Trajectory Classes" *Alcohol Clin Exp Res* 24 (6):882-91.

Nagin, Daniel. 2005. *Group-Based Modeling of Development*. Cambridge, MA: Harvard University Press.

Provonost P.J., M.R. Miller, R.M. Wachter. 2006. "Tracking Progress in Patient Safety: An Elusive Target." *Journal of the American Medical Association* 296 (6):696-99.

Raftery A.E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-64.

Romano, P.S., J.J. Geppert, S. Davies, M.R. Miller, A. Elixhauser, and K.M. McDonald. 2003. "A National Profile of Patient safety in U.S. Hospitals." *Health Affairs* 22 (2): 154-66.

Schneider, E.C. and A.M. Epstein. 1996. "Influence of Cardiac-Surgery Performance Reports on Referral Practices and Access to Care." *New England Journal of Medicine* 335(4): 251-256.

Schwartz, L.M., S. Woloshin, and J.D. Birkmeyer. 2005. "How Do Elderly Patients Decide Where to Go for Major Surgery? Telephone Interview Survey." *British Medical Journal* 331 (7520): 821-824.

Tay, A. 2003. "Assessing competition in hospital care markets: The importance of accounting for quality differentiation*." RAND Journal of Economics* 34 (4): 786-814.

Wasserman, Larry. 2004. *All of Statistics: A Concise Course in Statistical Inference*. New York, NY: Springer.

Werner, R.M. and D.A. Asch. 2005. "The Unintended Consequences of Publicly Reporting Quality Information." *Journal of the American Medical Association* 293 (10): 1239-1244.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Zhan, C. and M.R. Miller. 2003a. "Administrative Data Based Patient Safety Research: A Critical Review." *Qual Saf Health Care* 12 (supp II): ii58-63.

Zhan, C. and M.R. Miller. 2003b. "Excess Length of Stay, Charges, and Mortality Attributable to Medical Injuries During Hospitalization." *JAMA* 290 (14): 1868-74.

Zhan, C., E Kelley, H.P. Yang, M. Keyes, J. Battles, R.J. Borotkanics, and D. Stryer. 2005. "Assessing Patient Safety in the United States: Challenges and Opportunities." *Medical Care* 43 (3, suppl): I42-7.