



## Mortality Rate Estimation and Standardization for Public Reporting: Medicare's Hospital Compare

E. I. George, V. Ročková, P. R. Rosenbaum, V. A. Satopää & J. H. Silber

To cite this article: E. I. George, V. Ročková, P. R. Rosenbaum, V. A. Satopää & J. H. Silber (2017) Mortality Rate Estimation and Standardization for Public Reporting: Medicare's Hospital Compare, Journal of the American Statistical Association, 112:519, 933-947, DOI: [10.1080/01621459.2016.1276021](https://doi.org/10.1080/01621459.2016.1276021)

To link to this article: <https://doi.org/10.1080/01621459.2016.1276021>



View supplementary material [↗](#)



Accepted author version posted online: 20 Jan 2017.  
Published online: 20 Jan 2017.



Submit your article to this journal [↗](#)



Article views: 1045



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)



# Mortality Rate Estimation and Standardization for Public Reporting: Medicare's Hospital Compare

E. I. George<sup>a</sup>, V. Ročková<sup>b</sup>, P. R. Rosenbaum<sup>a</sup>, V. A. Satopää<sup>c</sup>, and J. H. Silber<sup>d</sup>

<sup>a</sup>Department of Statistics, University of Pennsylvania, Philadelphia, PA; <sup>b</sup>Department of Econometrics and Statistics at the Booth School of Business of the University of Chicago, Chicago, IL; <sup>c</sup>Department of Technology and Operations Management at INSEAD, Fontainebleau, France; <sup>d</sup>Department of Pediatrics and Anesthesiology & Critical Care, The University of Pennsylvania School of Medicine and Department of Health Care Management, The Wharton School, Philadelphia, PA

## ABSTRACT

Bayesian models are increasingly fit to large administrative datasets and then used to make individualized recommendations. In particular, Medicare's Hospital Compare webpage provides information to patients about specific hospital mortality rates for a heart attack or acute myocardial infarction (AMI). Hospital Compare's current recommendations are based on a random-effects logit model with a random hospital indicator and patient risk factors. Except for the largest hospitals, these individual recommendations or predictions are not checkable against data, because data from smaller hospitals are too limited to provide a meaningful check. Before individualized Bayesian recommendations, people derived general advice from empirical studies of many hospitals, for example, prefer hospitals of Type 1 to Type 2 because the risk is lower at Type 1 hospitals. Here, we calibrate these Bayesian recommendation systems by checking, out of sample, whether their predictions aggregate to give correct general advice derived from another sample. This process of calibrating individualized predictions against general empirical advice leads to substantial revisions in the Hospital Compare model for AMI mortality. To make appropriately calibrated predictions, our revised models incorporate information about hospital volume, nursing staff, medical residents, and the hospital's ability to perform cardiovascular procedures. For the ultimate purpose of comparisons, hospital mortality rates must be standardized to adjust for patient mix variation across hospitals. We find that indirect standardization, as currently used by Hospital Compare, fails to adequately control for differences in patient risk factors and systematically underestimates mortality rates at the low volume hospitals. To provide good control and correctly calibrated rates, we propose direct standardization instead. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received March 2015  
Revised November 2016

## KEYWORDS

Bayesian inference;  
Calibration by matching;  
Hierarchical random effects  
modeling; Individualized  
prediction; Predictive Bayes  
factors; Standardized  
mortality rates

## 1. Are Mortality Rates For AMI Lower at Some Hospitals?


### 1.1. Individualized Bayes Predictions Should Calibrate with Sound, Empirically Based General Advice

With a view to providing the public with information about the quality of hospitals, Medicare runs a website called "Hospital Compare" (<http://www.medicare.gov/hospitalcompare/>). Among other things, for each hospital, Hospital Compare provides information about the mortality rate of patients treated for a heart attack, or "acute myocardial infarction" (AMI). If you enter your zip code at the website, Hospital Compare will tell you about hospitals near where you live. Nationally, for a person who arrives at the hospital alive, the 30 day mortality rate following AMI is in the vicinity of 15%. The website's reported hospital-specific mortality rates are based on Medicare claims data and a random effects logit model in which hospitals enter as a random intercept and adjustments are made for risk factors describing individual patients, for instance, age and prior heart attacks. The number reported by Hospital Compare is essentially an indirectly standardized mortality rate for each hospital, adjusting

for measured risk factors describing the patient. An indirectly standardized rate is a constant multiple of a ratio of two predictions for the mortality of the patients actually treated at that hospital, namely, in the numerator, the model's predicted mortality rate if these patients were treated at this hospital, and in the denominator, the model's predicted mortality for the same patients if treated at what Hospital Compare considers to be a "typical" hospital. A ratio substantially above one is interpreted as "worse than average risk" and a ratio substantially below one is interpreted as "better than average risk." The website describes most hospitals as "no different than the national average."

Some small hospitals treat a few AMIs per year, whereas there is a hospital in New York that treats on average about two AMIs per day. Mortality rates from small hospitals are quite unstable, and the random intercepts model used by Hospital Compare shrinks these rates to resemble the National average. Their model says: "if there is not much data about your hospital, then we predict it to be average." For any one small hospital, there is not much data to contradict that prediction. So their model claims that the mortality rate at each small hospital is close to the National average. Is this a discovery or an assumption?

**CONTACT** Edward I. George  [edgeorge@wharton.upenn.edu](mailto:edgeorge@wharton.upenn.edu)  Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104.  
Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

© 2017 American Statistical Association

If it is a discovery, then it is a surprising discovery. A fairly consistent finding in health services research is that, adjusting for patient risk factors, mortality rates are typically higher at low volume hospitals (Luft, Hunt, and Maerki 1987; Halm, Lee, and Chassin 2002; Gandjour, Bannenberg, and Lauterbach 2003; Shahian and Normand 2003). Indeed, this pattern is unambiguously evident in the data used to fit the Hospital Compare model. Therefore, sound general advice would be to avoid low volume hospitals for treatment of AMI.

So, is the finding of average risk at small hospitals a discovery or an assumption? Actually, it is neither: it is a mistake. The model is not properly calibrated; see Dawid (1982) for discussion of calibration. Although there is very little data about any one small hospital, hence very little data to check a statement about one small hospital, there is plenty of data about small hospitals as a group. When Hospital Compare's predictions for all small hospitals are added up, it is unambiguously clear that the risk at small hospitals as a group is well above the national average; see Silber et al. (2010).

There is, here, a general principle. A Bayesian model can use all of the data to make an individualized prediction that is difficult to check as a single prediction. It is possible that this individualized prediction is better than relying upon general advice, because it is possible that this individualized prediction is tapping into distinctions evident in the data but not reflected in general advice. But if the general advice is correct as general advice, the individualized predictions should not aggregate to contradict it. As a check on whether a Bayesian model is calibrated, checking individualized predictions against general advice has two virtues. First, what it checks can fail to hold, so it can reject some models as inadequate. Second, what it checks is relevant: the check is against the advice you would fall back upon if individualized predictions were unavailable. A model may be detectably false in an irrelevant way—it may use a double exponential distribution where a logistic distribution would have been better—but that model failure may have negligible consequences for its recommendations. However, if the model contradicts correct general advice, then there is reason to worry about its individualized predictions. These general considerations are illustrated in Section 4.2.

## 1.2. Outline: Modeling, Calibrating, and Reporting Hospital Mortality Rates

In the current article, we show how the Hospital Compare model can be elaborated to yield improved predictions that no longer contradict general advice. We confirm such improvements by fitting the model in one sample and making predictions for another: in particular, we predict the outcome of the general advice that would be obtained from the second sample by an empirical study that made no use of the model. For the public reporting of these improved predictions, we propose a direct standardization approach that is effective at adjusting hospital mortality rate comparisons for patient mix differences between hospitals.

In Section 2, we apply a Bayesian implementation of the Hospital Compare model to recent Medicare data for AMI, obtaining results similar to those reported on the Hospital Compare web-page. Observing how it treats the various sources of

mortality rate variation, we then consider, in Section 3, whether the Hospital Compare model adequately describes the data. Specifically, in Section 3, we describe a sequence of hierarchical random effects logit models predicting AMI mortality from attributes of patients prior to admission, such as age, prior MI, and diabetes, and from the identity and attributes of individual hospitals, such as a hospital's volume, its capabilities in interventional cardiology and cardiac surgery, and the adequacy of its support staff in terms of nurses and residents. We also consider an interaction between a patient attribute and a hospital attribute, so the model becomes able to say that the best hospital for one patient may differ from the best hospital for another patient.

The models are evaluated in Section 4.1 on the basis of predictive Bayes factors that gauge their ability to make out-of-sample predictions. Then Section 4.2 checks whether the models are calibrated by (i) performing matched studies of general advice using out-of-sample data, studies that make no use of the model, and (ii) using the model's out-of-sample predictions to predict the results of those matched studies. In other words, the model's individualized predictions are aggregated to predict the results of a study that might have been used to generate general advice without individualized predictions. Some models are much better calibrated than others.

We turn to standardization for public reporting in Section 5. The parameters of a Bayesian model would be difficult for the public to understand. Hospital Compare reports indirectly standardized rates. We contrast and evaluate directly and indirectly standardized rates. We conclude that indirectly standardized rates should not be used for public reporting, but directly standardized rates work well. In Section 6, we describe what can be learned from our recommended approach to hospital mortality rate estimation. This includes mortality rate uncertainty intervals, classification of hospitals by low, average, and high mortality rates, and the influence of hospitals attributes on mortality. Section 7 concludes with a discussion. In supplemental appendices, we provide technical aspects of our Bayesian computational approach, as well as extensions and details of the various analyses considered throughout.

## 2. Hierarchical Bayesian Models for Adjusted Mortality Rates

### 2.1. The Data

Our data were obtained from Medicare records on  $N = 377,615$  AMI cases for patients admitted to  $H = 4289$  hospitals from July 1, 2009 to December 31, 2011. The first two years of data, up to June 30, 2011, were used to fit the models under consideration here and in Section 3. The remaining six months of data were then used for the model validations in Section 4. We will refer to these two datasets as the training data and the validation data, respectively.

Each case in our data contains an indicator of patient death within 30 days of admission, patient-specific demographics and risk factors (gender, age, history of diabetes, etc.), and hospital-specific attributes (volume, number of beds, etc.). We denote these variables as follows. For patient  $j$  in hospital  $h$ ,  $j = 1, 2, \dots, n_h$  and  $h = 1, 2, \dots, H$ , let  $y_{hj} \in \{0, 1\}$  be the binary

outcome for whether the patient died ( $y_{hj} = 1$ ) or did not die ( $y_{hj} = 0$ ) within 30 days of admission. Let  $\mathbf{x}_{hj}$  and  $\mathbf{z}_h$  be the accompanying vectors of patient attributes and hospital attributes, respectively. The number of patients per hospital  $n_h$  varied a great deal over our data, ranging from 1 to 2782 patients, with a median value of 79. The three year volume of all Medicare AMI admissions at hospital  $h$ , which we denote as  $\text{vol}_h$ , is a particular characteristic that will turn out to figure prominently in our modeling of mortality rates throughout.

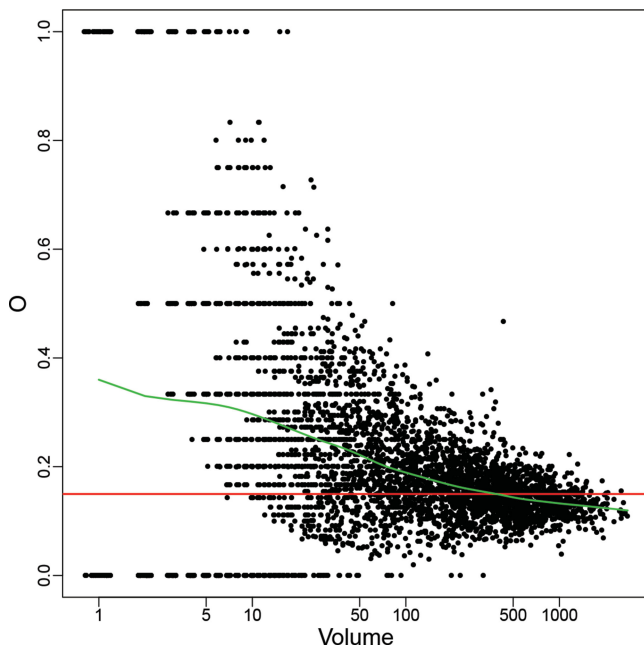
## 2.2. The Hospital Compare Random Effects Model

To motivate the modeling of mortality rates for our Medicare data, let us begin with Figure 1, a plot of the raw observed hospital mortality rates  $O_h = \frac{1}{n_h} \sum_j y_{hj}$  by volume  $\text{vol}_h$ . As would be expected,  $O_h$  variation is largest at low volume hospitals where  $n_h$  is small, and then steadily decreases as hospital volume increases. Also evident in the plot is a steadily decreasing average mortality rate, summarized by a smoothing spline, which is highest at low volume hospitals. This spline crosses the overall average patient mortality rate line  $\bar{y} = 0.1498$  at a hospital volume of about 450. An issue of central interest is the extent to which this average patient mortality rate curve can be explained by patient attributes and/or hospital characteristics.

Recognizing patient and hospital effects as potential sources of mortality rate variation, Medicare's Hospital Compare (Yale New Haven Health Services Corporation 2014, Appendix 7 A) uses a random-effects logit model to estimate underlying hospital mortality rates. Proposed by Krumholz et al. (2006) for this context, this model is of the form

$$Y_{hj} | \alpha_h, \beta, \mathbf{x}_{hj} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p_{hj}) \text{ where } \text{logit}(p_{hj}) = \alpha_h + \mathbf{x}'_{hj}\beta \quad (1)$$

$$\alpha_h | \mu_\alpha, \sigma_\alpha^2 \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_\alpha, \sigma_\alpha^2). \quad (2)$$



**Figure 1.** Raw observed hospital mortality rates  $O_h$  by  $\text{vol}_h$ . Overall average rate indicated by the red horizontal line. Average rate by  $\text{vol}_h$  summarized by the green superimposed smoothing spline.

Here,  $P(Y_{hj} = 1) = p_{hj} = \text{logit}^{-1}(\alpha_h + \mathbf{x}'_{hj}\beta)$  is the  $h$ th patient's underlying 30-day mortality rate, which is determined by a hospital effect  $\alpha_h$  and a patient effect  $\mathbf{x}'_{hj}\beta$ . The hospital effects  $\alpha_h$  are modeled as independent normal random effects drawn from a single normal distribution, which does not depend on any hospital attributes. On the other hand, the patient effects  $\mathbf{x}'_{hj}\beta$ , which explicitly depend on patient attributes  $\mathbf{x}_{hj}$ , are transmitted through a common fixed effect vector  $\beta$ . Under this model, the underlying average 30-day mortality rate for patients treated at hospital  $h$  is given by

$$P_h = \frac{1}{n_h} \sum_{j=1}^{n_h} p_{hj}. \quad (3)$$

To mesh with our labeling of models proposed later in Section 3.2, we will refer to the Hospital Compare model (1)–(2) as the (C,C) model because it constrains both the mean and the variance of the  $\alpha_h$  distribution to be constant.

For the implementation of this (C,C) model, we propose a fully Bayesian approach with the relatively noninfluential, neutral conjugate priors

$$\beta | \sigma_\beta^2 \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_d), \quad \sigma_\beta^2 \sim \text{IG}(1, 1) \quad (4)$$

for the fixed effects parameters in (1), and

$$\mu_\alpha \sim \mathcal{N}(0, g\sigma_\alpha^2), \quad g \sim \text{IG}(1, 1), \quad \sigma_\alpha^2 \sim \text{IG}(1, 1) \quad (5)$$

for the hyperparameters of the random effect distribution in (2). For compatibility with the model elaborations proposed in Section 3, we have used a heavy tailed conjugate hyper  $g$ -prior for  $\mu_\alpha$ .

With such a fully Bayes model, all inferences about mortality rates, hospital effects, patient effects, and functions of these, can be obtained from the posterior distributions  $\pi(\mathbf{p} | \mathbf{y})$ ,  $\pi(\alpha | \mathbf{y})$  and  $\pi(\beta | \mathbf{y})$ , where  $\mathbf{p}$ ,  $\alpha$ ,  $\beta$ , and  $\mathbf{y}$  are the complete vectors of mortality rates, hospital effects, fixed effects, and observed mortality indicators, respectively. In particular, we use posterior means  $\hat{\mathbf{p}}$ ,  $\hat{\alpha}$ , and  $\hat{\beta}$  as estimates throughout, with 95% posterior density intervals to describe their uncertainty. As described in Appendix A.1, these can all be efficiently computed by Markov chain Monte Carlo (MCMC) posterior simulation using a Pólya-Gamma latent variable augmentation (Polson, Scott, and Windle 2013). As would be expected under heavy tailed priors such as (4) and (5), our posterior mean estimates from the (C,C) model are very similar to the constrained likelihood estimates used by Hospital Compare with the SAS 9.3 GLIMMIX software. Indeed, the  $\hat{\alpha}_h$  GLIMMIX estimates and the  $\hat{\alpha}_h$  Bayes estimates here had a correlation of 0.9994.

Figure 2(a) plots the  $\hat{P}_h$  posterior mean estimates by  $\text{vol}_h$  for the (C,C) model. We see immediately that both the  $O_h$  values and the decreasing average mortality rate spline from Figure 1 have been shrunk toward the overall mean mortality rate line  $\bar{y} = 0.1498$ . As we would expect, the  $O_h$  realizations have simply added extra variability to their underlying  $P_h$  values. Nevertheless, it appears that substantial  $\hat{P}_h$  variability remains, especially at the small volume hospitals where  $n_h$  in (3) is small.

Insight into the source of the  $\hat{P}_h$  variability under the (C,C) model is obtained from Figure 3(a), a plot of the  $\hat{\alpha}_h$  posterior



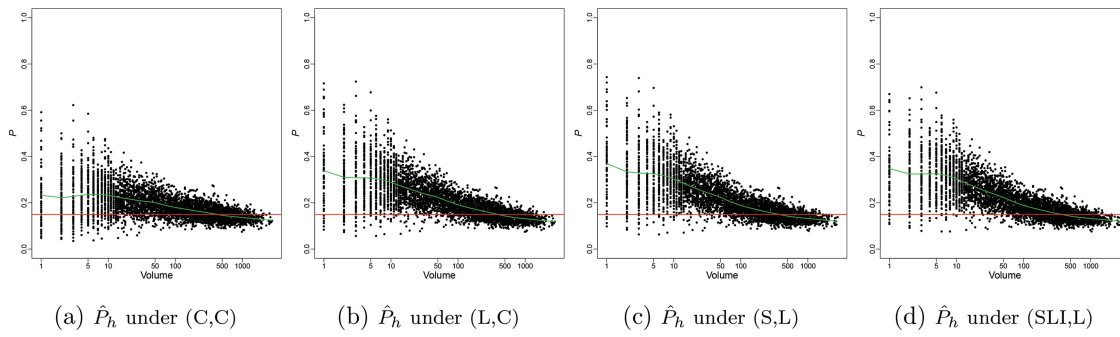


Figure 2.  $\hat{P}_h$  versus  $\text{vol}_h$ .

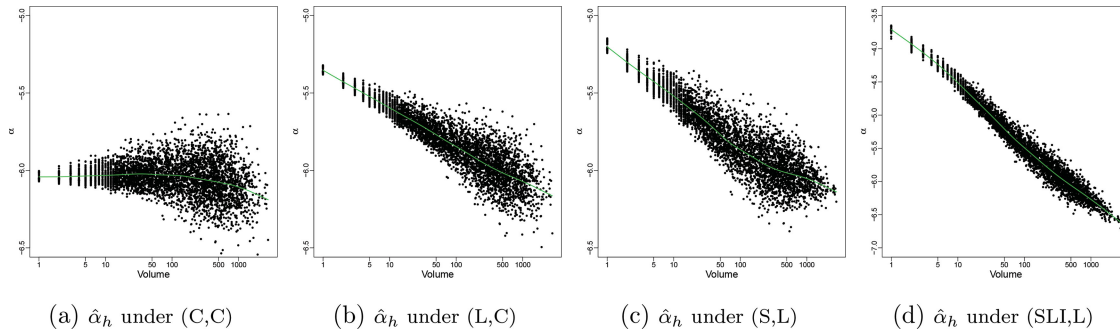


Figure 3.  $\hat{\alpha}_h$  versus  $\text{vol}_h$ .

mean estimates by  $\text{vol}_h$ . Whereas the  $\hat{P}_h$  manifest larger variation at the low volume hospitals, as well as an elevated and decreasing average mortality rate, the  $\hat{\alpha}_h$  manifest exactly the opposite behavior (though on a different scale than the  $\hat{P}_h$ ). The variation of the  $\hat{\alpha}_h$  is smallest at the low volume hospitals, where the average  $\hat{\alpha}_h$  smoothing spline is flat and unrelated to hospital volume. Since  $\alpha_h$  and  $\mathbf{x}'_{hj}\boldsymbol{\beta}$  are the only components of  $p_{hj}$  in (1), the variation of the (C,C) model  $\hat{P}_h$  at the low volume hospitals is being driven almost entirely by the variation of patient effects  $\mathbf{x}'_{hj}\boldsymbol{\beta}$ ,  $j = 1, \dots, n_h$  across hospitals. Thus, under the (C,C) model, the elevated average mortality rates at the lower volume hospitals is coming from a riskier patient case-mix distribution at those hospitals rather than from hospital effect differences. A primary purpose of the Hospital Compare analysis is to adjust for patient case-mix variation with an indirect standardization of the hospital mortality estimates. This standardization effectively eliminates all mortality rate differences between the low volume hospitals as will be seen in Figure 4(a) in Section 5. As discussed in Section 1, such a conclusion is at odds with the general finding in the literature that patient risk-adjusted mortality rates are typically higher at low volume hospitals (Luft, Hunt, and Maerki 1987; Halm, Lee, and Chassin 2002; Gandjour, Bannenberg, and Lauterbach 2003; Shahian and Normand 2003).

### 3. Hierarchical Modeling of the Random Hospital Effects

#### 3.1. The Illusion of Safe Shrinkage Estimation

The absence of an elevated level of low volume hospital effects in Figure 3(a) turns out to be an artifact of the (C,C) model, as

will be seen by comparison with alternative models. In leaving hospital characteristics out of the model and treating the  $\alpha_h$ 's as independent of volume  $\text{vol}_h$ , the (C,C) model has not allowed the data to speak to this possibility. Indeed, the pattern in Figure 3(a) is consistent with the strong random effects assumption (2) of normally distributed  $\alpha_h$ 's with a common mean  $\mu_\alpha$  and variance  $\sigma_\alpha^2$ . Under such a normal prior, we would expect all the  $\hat{\alpha}_h$  estimates to be shrunk toward a single common mean. Such shrinkage would be especially pronounced for those  $\hat{\alpha}_h$  for which there is less sample information, namely, the  $\hat{\alpha}_h$ 's for the low volume hospitals. This is exactly what we see.

Although shrinkage estimation has the potential to improve noisy estimates, such as the raw small hospital mortality rates in our setting, it can only do so if the shrinkage targets, here the means of the  $\alpha_h$ 's, are appropriately specified. Contrary to the commonly held belief that shrinkage estimation can do no harm, which can be the case in certain stylized contexts, shrinkage estimation with a model that is at odds with the data can be very detrimental. With an unforgiving, nonrobust normal prior that imposes strong shrinkage, the resulting estimates will be poor and misleading if shrinkage targets are incorrectly specified; see Berger (1985).

The evident and plausible relationship between mortality rates and volume suggests that it would be more reasonable to shrink mortality rates toward the mean rates of hospitals with similar volumes. Unfortunately, this does not happen with the random effects distribution (2), which shrinks all rates toward a single overall rate.

#### 3.2. Modeling the Random Hospital Effects

The key to the development of a better hierarchical Bayes model for our data is to elaborate the random effects distribution (2)

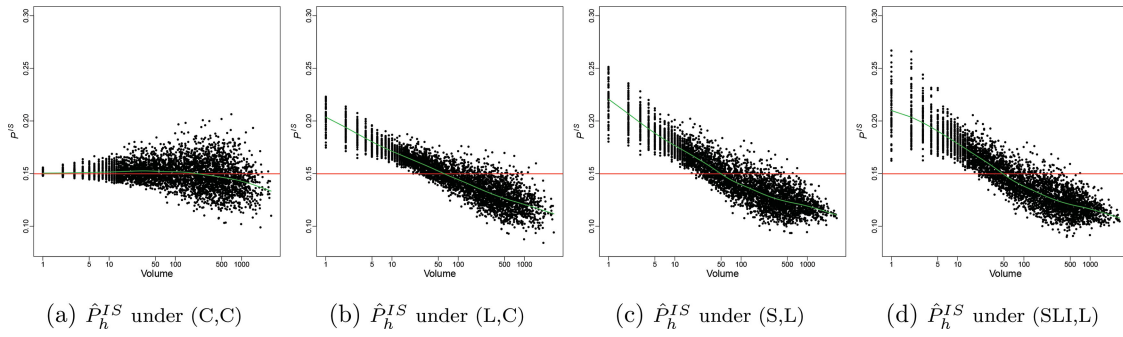


Figure 4.  $\hat{P}_h^{IS}$  versus  $v o l_h$ .

in a way that will allow the data to inform us of any potential relationship between hospital mortality rates and volume and hospital attributes such as volume. For this purpose, we propose hierarchical logit model formulations of the form

$$Y_{hj} | \alpha_h, \beta, \mathbf{x}_{hj} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p_{hj}) \text{ where } \text{logit}(p_{hj}) = \alpha_h + \mathbf{x}_{hj}'\beta \quad (6)$$

$$\alpha_h | \mu_h(\mathbf{z}), \sigma_h^2(\mathbf{z}) \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_h(\mathbf{z}), \sigma_h^2(\mathbf{z})), \quad (7)$$

where  $p_{hj} = P(Y_{hj} = 1)$  is the  $h$ th patient's underlying 30-day mortality rate. As in (1),  $\text{logit}(p_{hj})$  in (6) is still the sum of a hospital effect  $\alpha_h$  plus a fixed effect  $\mathbf{x}_{hj}'\beta$ . However, (7) now allows the mean  $\mu_h(\mathbf{z})$  and the variance  $\sigma_h^2(\mathbf{z})$  of the hospital effects distribution to be functions of the hospital attributes  $\mathbf{z}$ . As before, the fixed effects  $\mathbf{x}_{hj}'\beta$  in (6) still account for patient risk variation via the patient attributes  $\mathbf{x}_{hj}$ , but the random effects distribution  $\mathcal{N}(\mu_h(\mathbf{z}), \sigma_h^2(\mathbf{z}))$  in (7) now allows the  $\alpha_h$ 's to more fully account for hospital-to-hospital variation via the hospital attributes  $\mathbf{z}$ . Note that as a matter of convenience, we have separated the roles of  $\mathbf{x}_{hj}$  in (6) and  $\mathbf{z}$  in (7) to be consistent with their variation at the patient and hospital levels, respectively. Because hospital attributes  $\mathbf{z}$  do not vary within hospitals, this partition avoids their inclusion within the patient level linear component (6), where modeling their effects would be complicated by the collinearity of the intercept estimates with hospital attributes deployed at the patient level.

We now proceed to consider specific formulations for  $\mu_h(\mathbf{z})$  and  $\sigma_h^2(\mathbf{z})$  that yield better calibrated predictions for our data. As noted earlier, we refer to the Hospital Compare (1)–(2) specification of (6)–(7), namely, the one for which  $\mu_h(\mathbf{z}) \equiv \mu_\alpha$  and  $\sigma_h^2(\mathbf{z}) \equiv \sigma_\alpha^2$ , as the Constant-Constant (C,C) model. Each of the formulations proposed below will be a relaxation that nests the (C,C) model as a special case, thereby allowing the data to ignore the hospital attributes if they do not lead to better predictions. Thus, these formulations will let the data speak rather than override what the data have to say.

### 3.3. Modeling $\alpha_h$ as a Function of Volume

To shed light on the issue of whether hospital mortality rates are related to volume after accounting for patient mix effects, we begin by considering formulations for the mean and variance of the  $\alpha_h$  hospital effects in (7) as functions of the hospital attribute volume  $v o l_h$  only. We begin with a simple linear specification, and then proceed to consider a more flexible model

that allows for a more refined description of the underlying relationship. This flexible formulation will then serve as a foundation for the further addition of hospital covariates and interactions to our final model in Section 3.4.

Before proceeding, we should emphasize that, although the application of our models will be seen to unambiguously reveal a strong association between high hospital mortality rates and low volume hospitals, we are not addressing the issue of whether this relationship is causal. Our goal is mainly to confirm the aforementioned finding of such an association in the literature, and to show that by including hospital volume in our models, we get better and more informative predictions with the Medicare data. Such predictions will help guide patients toward safer hospitals.

To provide further insight into the relationship between mortality rates and hospital size, we also examined the relationship between mortality rates and `beds2008`, the number of beds in 2008, a hospital attribute that is indisputably exogenous to our observed mortality rates. As shown in Appendix A.3 of the supplemental material, a strong association between these two variables persists.

#### 3.3.1. A Simple Linear Emancipation of the Means

We begin with perhaps the simplest elaboration of the mean and variance functions for (7),

$$\mu_h(\mathbf{z}) = \gamma_0 + \gamma_1 \log(v o l_h + 1), \quad \sigma_h^2(\mathbf{z}) \equiv \sigma_\alpha^2, \quad (8)$$

a linear relaxation of the mean that keeps the variance constant. For the full model (6)–(7) with this specification, we add the conjugate fixed effect prior (4) for  $\beta$ , and add the conjugate hyperparameter priors

$$(\gamma_0, \gamma_1)' | g, \sigma_\alpha^2 \sim \mathcal{N}_2(\mathbf{0}, g\sigma_\alpha^2 \mathbf{I}_2), \quad g \sim \mathcal{IG}(1, 1), \quad \sigma_\alpha^2 \sim \mathcal{IG}(1, 1), \quad (9)$$

where  $g$  and  $\sigma_\alpha^2$  are a priori independent. We will refer to the hierarchical logit model formulation with this Linear-Constant specification as the (L,C) model. Note that the (L,C) model nests the (C,C) model as the special case for which  $\gamma_1 = 0$ .

Application of the (L,C) model to our data produced the hospital mortality rate and hospital effect estimates  $\hat{P}_h$  and  $\hat{\alpha}_h$  displayed in Figures 2(b) and 3(b). We see immediately that compared to their (C,C) counterparts in Figure 2(a), the (L,C)  $\hat{P}_h$ 's are generally higher at the lower volume hospitals where the average rate smoothing spline has been raised substantially. This is evidently a consequence of their component  $\hat{\alpha}_h$ 's in Figure 3(b), which are dramatically different from their (C,C) counterparts

in Figure 3(a). The (L,C)  $\hat{\alpha}_h$ 's are now substantially higher at the low volume hospitals, with a clear downward sloping linear trend in  $\log(\text{vol}_h + 1)$  summarized by the superimposed smoothing spline. With a posterior mean estimate  $\hat{\gamma}_1 = -0.106$  and a 95% credible interval  $(-0.116, -0.096)$ , the data have expressed an unambiguous preference for a downward sloping (L,C) mean specification (8) over the (C,C) constant specification (2) for which  $\gamma_1 = 0$ . As will be seen with the more formal predictive model comparisons in Section 4.1, the (L,C) model  $\hat{P}_h$  improve very substantially over their (C,C) counterparts.

### 3.3.2. A Spline-Log-Linear Emancipation of the Means and Variances

With only a simple linear relaxation of the mean function, the (L,C) model released volume to reveal higher mortality rates explained by dramatically higher hospital effects at the low volume hospitals. To further release the explanatory power of volume, we consider a more flexible relaxation of both the mean and variance functions. In particular, we consider a spline specification for  $\mu_h(\mathbf{z})$  coupled with a log-linear specification for  $\sigma_h^2(\mathbf{z})$ .

For the spline mean specification, let  $\mathbf{v} = (\log(\text{vol}_1 + 1), \dots, \log(\text{vol}_H + 1))'$  be the vector of all hospital specific volumes. We construct a B-spline basis of degree  $d$  and number of knots  $\kappa$ , represented by the columns of  $\mathbf{B}_{d,\kappa}(\mathbf{v})$ , an  $H \times k$ , ( $k = (d + 1) + \kappa$ ) matrix. Letting  $\mathbf{b}_h(\mathbf{v})$  be the  $h$ th row of  $\mathbf{B}_{d,\kappa}(\mathbf{v})$ , our spline specification is obtained as

$$\mu_h(\mathbf{z}) = \mathbf{b}_h(\mathbf{v})\boldsymbol{\gamma}_S, \quad (10)$$

where  $\boldsymbol{\gamma}_S$  is a  $k \times 1$  vector of spline regression coefficients. To this we add a prior on  $\boldsymbol{\gamma}_S$  of the form

$$\boldsymbol{\gamma}_S | g_S, \sigma_\alpha^2 \sim \mathcal{N}_k(\mathbf{0}, g_S \sigma_\alpha^2 \mathbf{P}^{-1}), \quad g_S \sim \mathcal{IG}(1, 1), \quad (11)$$

where  $\mathbf{P}$  is a banded matrix that penalizes second-order differences between adjacent spline coefficients, and  $g_S$  serves as a roughness penalty that determines the wiggleness of the resulting curve. (The usual spline roughness penalty is  $\lambda = 1/g_S$ .) With this penalization, a nested linear parametric form is obtained as  $g_S \rightarrow 0$ . Note that the conjugate inverse-gamma prior on  $g_S$  is often considered in the context of hyper- $g$ -priors for variable selection. This P-spline (penalized B-spline) formulation allows us to begin with a rich B-spline basis  $\mathbf{B}_{d,\kappa}(\mathbf{v})$  with many knots, and use regularization to circumvent the difficulties of optimizing the number and position of knots.

For the log-linear variance specification (Box and Meyer 1986; Gu et al. 2009), we set

$$\sigma_h^2(\mathbf{z}) = \exp\{\delta \text{vol}_h\} \sigma_\alpha^2, \quad (12)$$

which nests the previous case  $\sigma_h^2(\mathbf{z}) \equiv \sigma_\alpha^2$  when  $\delta = 0$ . To this we add the prior

$$\delta | \sigma_\alpha^2 \sim \mathcal{N}(0, g_\delta \sigma_\alpha^2), \quad g_\delta \sim \mathcal{IG}(1, 1), \quad (13)$$

and then complete the entire Bayesian specification of (10)–(13) with

$$\sigma_\alpha^2 \sim \mathcal{IG}(1, 1), \quad (14)$$

where  $g_S$ ,  $g_\delta$ , and  $\sigma_\alpha^2$  are all assumed a priori independent. We will refer to the hierarchical logit model formulation with this spline-log-linear specification as the (S,L) model.

Application of this (S,L) model with a B-spline basis of degree  $d = 3$  and  $\kappa = 17$  knots yields the hospital rate and effect estimates  $\hat{P}_h$  and  $\hat{\alpha}_h$  in Figures 2(c) and 3(c). These estimates again support the story revealed by the (L,C) model that low volume hospitals have generally higher mortality rates driven by substantially higher hospital effects. Having been freed from the constraints of linearity and constant variance, the low volume  $\hat{\alpha}_h$  estimates are now even higher and more dispersed than their (C,C) counterparts in Figure 3(b), manifesting a decreasing nonlinear trend. And with a posterior mean estimate of  $\hat{\delta} = -0.00112$  and a 95% credible interval of  $(-0.00160, -0.00066)$ , the data have also expressed a preference for the log linear variance specification (12) as well. Although the (L,C) and (S,L)  $\hat{P}_h$  plots in Figure 2(b) and 2(c) seem very similar, manifesting considerable patient-mix variability, the formal predictive comparisons that we will see in Section 4.1 confirm the (S,L) model estimates as a further improvement.

### 3.4. Adding Further Hospital Attributes and Interactions to the Model

With our (S,L) model as a foundation, we now consider enlarging the model to incorporate further hospital attributes. This is most simply done by adding them as linear terms to  $\mu_h(\mathbf{z})$  in (10). This yields the spline-linear mean specification

$$\mu_h(\mathbf{z}) = \mathbf{b}_h(\mathbf{v})\boldsymbol{\gamma}_S + \mathbf{z}_h'\boldsymbol{\gamma}_L, \quad (15)$$

where  $\mathbf{z}_h$  is an  $r \times 1$  vector of hospital  $h$  attributes and  $\boldsymbol{\gamma}_L$  is an  $r \times 1$  vector of linear regression coefficients. This form is then completed with the priors

$$\boldsymbol{\gamma}_S | g_S, \sigma_\alpha^2 \sim \mathcal{N}_k(\mathbf{0}, g_S \sigma_\alpha^2 \mathbf{P}^{-1}), \quad g_S \sim \mathcal{IG}(1, 1), \quad (16)$$

$$\boldsymbol{\gamma}_L | g_L, \sigma_\alpha^2 \sim \mathcal{N}_r(\mathbf{0}, g_L \sigma_\alpha^2 \mathbf{I}_r), \quad g_L \sim \mathcal{IG}(1, 1), \quad (17)$$

where  $g_S$  and  $g_L$  are priori independent of each other and of  $\sigma_\alpha^2 \sim \mathcal{IG}(1, 1)$ .

Going further, hospital attributes can also be incorporated as patient-hospital interactions of the form  $x_{hj} * z_h$ , products of particular attributes in  $\mathbf{x}_{hj}$  and  $\mathbf{z}_h$ , respectively. Because the values of such interaction terms vary at the patient level, these would be added as covariates to the linear fixed effects part of the model. Without such interactions, the model would say that one hospital,  $h$ , is either better or worse than another,  $h'$ , for every patient. Of course, there is no reason to restrict attention to models with this feature and no reason to expect the world to be well described by such a model. Patient-hospital interactions remove this limitation.

Keeping the log-linear variance formulation (12)–(13), we shall refer to the hierarchical logit formulation with this spline-linear-interaction specification as the (SLI,L) model. Applying an instance of this model to our data, we added three hospital attributes named NTBR, RTBR, and PCI as linear terms in (15). The nurse-to-bed-ratio (NTBR) and the resident-to-bed-ratio (RTBR) are continuous hospital variables that describe the density of support staff at a hospital. The binary hospital variable PCI is a catch-all for the ability of a hospital to perform

any of the following procedures: percutaneous coronary interventions such as percutaneous transluminal coronary angioplasty (PTCA), a stent, or a coronary artery bypass graft (CABG) surgery. The ability to perform these procedures is common in large volume hospitals and much less common in small volume hospitals.

We also added a patient-hospital interaction,  $\text{age}_{hj} * \log(\text{vol}_h + 1)$ , appending it to  $\mathbf{x}_h$  as an additional covariate for the fixed effects term  $\mathbf{x}_{hj}'\boldsymbol{\beta}$ . With this interaction, the model may provide mortality rate estimates that favor one hospital for a younger Medicare patient, say aged 68, and a different hospital for another older Medicare patient, say aged 90.

Application of this particular (SLI,L) model to our data produced the hospital rate and effect estimates  $\hat{P}_h$  and  $\hat{\alpha}_h$  displayed in Figures 2(d) and 3(d). These estimates again convey the same message as the (L,C) and (S,L) estimates in Figures 2(b) and 2(c) and 3(b) and 3(c), namely, that low volume hospital mortality rates are generally higher, driven by low volume hospital  $\hat{\alpha}_h$ 's, which exhibit a clear decreasing average trend as  $\text{vol}_h$  increases. However, the  $\hat{\alpha}_h$  in Figure 3(d) now vary more from high to low (note the changed vertical scale), a consequence of adding the  $\text{age}_{hj} * \log(\text{vol}_h + 1)$  interaction, which has served to model a portion of the hospital effect variation at the patient level. With a patient-hospital interaction in the model, the  $\hat{\alpha}_h$ 's no longer capture the entirety of the hospital effects. And once again, although the  $\hat{P}_h$ 's for the three models in Figures 2(b)–2(d) look very similar, model comparisons in Section 4.1 will show that the (SLI,L) model leads to still further predictive improvements.

### 3.5. Further Potential Elaborations

As will be confirmed in Section 4, our (L,C), (S,L), and (SLI,L) models have served to reveal the inadequacies of the (C,C) model. Paving the way for further improvements, these hierarchical elaborations of the random effects model are hardly the end of the story. Indeed, it is clear that many further elaborations may be promising, for example, by adding more hospital attributes as linear terms, spline terms, or patient-hospital interactions. One might also consider elaborations of the log-linear variance specification (12) that include more hospital attributes, for example,  $\sigma_h^2(\mathbf{z}) = \exp\{\mathbf{z}'_h \boldsymbol{\delta}\} \sigma_\alpha^2$ , where  $\mathbf{z}_h$  is a  $q \times 1$  vector of hospital  $h$  characteristics (possibly different from  $\mathbf{x}_h$  above) and  $\boldsymbol{\delta}$  is a  $q \times 1$  regression vector.

Going further, one could also consider different families for random effects distributions. More robust parametric distributions such as the Cauchy or  $t$ -distributions would serve to down-weight the influence of extremes. Even more flexibility could be obtained with nonparametric prior distributions. Indeed, Guglielmi et al. (2014) proposed modeling hospital coronary mortality rates with a Bayesian hierarchical logit model analogous to our (L,C) model but with a dependent Dirichlet process for the random effects. Such an elaboration opened the door for clustering hospitals into groups with identical mortality rates. Other interesting nonparametric random effect logit models that also incorporate hospital process indicators for modeling and clustering hospital coronary mortality rates were proposed by Grieco, Ieva, and Paganoni (2012). Such nonparametric elaborations provide promising new routes for improved mortality rate modeling.

A model for hospital mortality rates can be used for a variety of purposes, not just public reporting. Spiegelhalter et al. (2012) discussed the issues that arise in different applications of such models.

## 4. Model Evaluation

### 4.1. Predictive Bayes Factor Model Comparisons

Following the traditional Bayesian model choice formalism, we use Bayes factors to compare the performance of the proposed models (L,C), (S,L), and (SLI,L) with the performance of the Hospital Compare (C,C) model. For this purpose, we turn to out-of-sample predictive Bayes factors rather than in-sample Bayes factors. As is well recognized, in-sample Bayes factors based on diffuse parameter priors, such as those we have used with our training data, are unreliable criteria for model comparisons (Cox 1961; Berger 2006). Furthermore, because prediction is the intended use of these models, comparisons based on out-of-sample performance are of fundamental relevance here. Thus, we use predictive Bayes factors evaluated on the validation data using posterior rather than prior predictive likelihoods (Gelfand and Dey 1994).

Posterior predictive likelihoods are obtained by averaging the probability of the validation data with respect to a data-updated “prior” distribution using the training data. Here, the predictive likelihood for model  $\mathcal{M}$  is obtained as

$$\pi(\mathbf{y}_{\text{val}} | \mathbf{y}_{\text{tr}}, \mathcal{M}) = \int_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \pi(\mathbf{y}_{\text{val}} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{M}) \pi(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{tr}}, \mathcal{M}) d\boldsymbol{\alpha} d\boldsymbol{\beta}, \quad (18)$$

where  $\mathbf{y}_{\text{val}}$  and  $\mathbf{y}_{\text{tr}}$  are the validation data and training data  $\mathbf{y}$  values, respectively, and  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_H)'$ . Note that the training data posterior  $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{tr}}, \mathcal{M})$  now serves as a stable and non-diffuse prior for the validation data. The predictive Bayes factor for comparison of model  $\mathcal{M}_1$  versus  $\mathcal{M}_2$  is then naturally defined as the ratio of the two posterior predictive likelihoods (Gelfand and Dey 1994; Kass and Raftery 1995),

$$\text{BF}_{\mathcal{M}_1, \mathcal{M}_2} = \frac{\pi(\mathbf{y}_{\text{val}} | \mathbf{y}_{\text{tr}}, \mathcal{M}_1)}{\pi(\mathbf{y}_{\text{val}} | \mathbf{y}_{\text{tr}}, \mathcal{M}_2)}.$$

Evaluation of the predictive Bayes factor is obtained by Monte Carlo integration of the posterior predictive Bayes likelihoods using posterior parameter samples from the MCMC output, as described in Section A.1, based on the training data. Using the simulated values  $\boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(s)} \sim \pi(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{tr}}, \mathcal{M}_h)$  from (A.4), these approximations to the posterior predictive likelihoods are obtained by the empirical averages

$$\hat{\pi}(\mathbf{y}_{\text{new}} | \mathbf{y}_{\text{tr}}, \mathcal{M}_i) = \frac{1}{M} \sum_{s=1}^S \pi(\mathbf{y}_{\text{val}} | \boldsymbol{\alpha}^{(s)}, \boldsymbol{\beta}^{(s)}, \mathcal{M}_i), \quad i = 1, 2.$$

The log posterior predictive Bayes factors comparing each of the (L,C), (S,L), (SLI,L) models with the (C,C) model are reported in Table 1. The predictive improvement over the (C,C) model by every one of our models is very large. Beginning with the (L,C) model, which simply allowed hospital effect means to



**Table 1.** Out-of-sample log posterior predictive Bayes factor comparisons to the (C,C) model.

Model	(L,C)	(S,L)	(SLI,L)
	27.54	32.13	37.96

be linear in volume rather than constant, the predictive likelihood increased by a huge factor of  $e^{27.54}$ . Each subsequent elaboration led to a further increase—moving from linear to spline-log-linear in volume (S,L), adding three hospital covariates and a patient-volume interaction—culminating in a predictive likelihood increase of  $e^{37.96}$  for our (SLI,L) model, which was the very best.

#### 4.2. Out-of-Sample Calibration of Aggregated Individual Predictions Against Empirical Studies of General Advice

The Bayesian model predicts future mortality rates at individual hospitals. For many hospitals, there are too few AMI patients to permit a serious test of the model's predictions at that hospital. Here, we calibrate the model by comparing its predictions to the general advice people would otherwise fall back on if individualized predictions were not available. Specifically, we conduct an out-of-sample observational study checking the general advice, then, we determine which models predict the results of that observational study with reasonable accuracy.

To illustrate, we consider the advice that one should avoid hospitals that rarely treat AMI. As noted earlier, the literature strongly suggests this is good general advice, although it is difficult to know whether it is good advice for any single hospital that treats few AMIs—after all, such a hospital provides few patients upon which to base a mortality rate.

Using the validation sample that was not used to build the model, we look at the 20% of hospitals with the lowest volume. This consisted of 747 low volume (LV) hospitals each with  $\text{vol}_h \leq 23$  AMI in Medicare patients over three years, that is, on average, at most one AMI patient in Medicare every 1.57 months. In the six-month Medicare validation sample, there were a total of 1353 AMI patients at such hospitals. Each such patient was matched to five patients at the 20% hospitals with the highest Medicare volume, defined to be the 753 hospitals with  $\text{vol}_h \geq 467$  over three years, or at least one Medicare AMI patient every 2.34 days. In a conventional way (Rosenbaum 2010; Stuart 2010), the matching combined some exact matching, a caliper on the propensity score, and optimal matching based on a Mahalanobis distance. Here, the propensity score predicted the low or high volume categories using a logit model and the covariates in Table 2. The training sample's estimate of risk of death was used as an out-of-sample risk or prognostic score in the validation sample, as suggested by Hansen (2008). Hansen's (2007) `optmatch` package in R was used.

Before discussing the results of this comparison, a few words of caution are needed. In every observational study, there is

**Table 2.** Covariate balance before and after matching. The table compares all 1302 patients at low volume hospitals to all 50,278 patients at high volume hospitals (All) and to 6510 high-volume controls matched 5-to-1 (Matched). The matching controlled the listed covariates that described the condition of the patient prior to admission. Standardized differences are differences in means in units of a pooled standard deviation prior to matching.

Patient covariates	Covariate means			Standardized differences	
	Low volume	High volume		Before matching	After matching
		Matched	All		
Number of patients	1302	6510	50,278	1302 versus 50,278	1302 versus 6510
Prior PTCA	0.06	0.06	0.16	−0.34	−0.02
Prior CABG	0.08	0.09	0.10	−0.07	−0.03
Heart failure	0.21	0.23	0.14	0.19	−0.04
Prior MI	0.12	0.12	0.09	0.10	−0.01
Anterolateral MI	0.05	0.05	0.10	−0.20	0.00
Inferolateral MI	0.07	0.07	0.15	−0.27	0.00
Unstable angina	0.03	0.03	0.02	0.03	0.01
Chronic athero.	0.46	0.47	0.82	−0.80	−0.01
CPR failure shock	0.05	0.06	0.06	−0.01	−0.02
Valvular heart dis.	0.12	0.13	0.20	−0.21	−0.02
Hypertension	0.66	0.67	0.71	−0.13	−0.04
Stroke	0.02	0.02	0.01	0.04	0.00
Cerebrovasc.	0.06	0.06	0.04	0.07	−0.01
Renal failure	0.14	0.16	0.12	0.08	−0.05
COPD	0.24	0.24	0.21	0.08	−0.01
Pneumonia	0.24	0.23	0.12	0.32	0.03
Diabetes	0.34	0.35	0.36	−0.04	−0.02
Malnutrition	0.04	0.04	0.05	−0.03	0.00
Dementia	0.22	0.23	0.12	0.29	−0.02
Paraplegia	0.03	0.03	0.02	0.06	−0.00
Peripheral vas. dis.	0.06	0.07	0.06	0.01	−0.01
Cancer	0.03	0.03	0.02	0.03	0.00
Trauma	0.12	0.11	0.10	0.04	0.01
Psych.	0.04	0.03	0.02	0.11	0.07
Chronic liver dis.	0.01	0.01	0.01	−0.03	−0.00
Male	0.41	0.40	0.54	−0.27	0.02
Age (years)	84.3	84.2	77.7	0.80	0.01
logit(Propensity score)	−2.94	−2.97	−4.34	1.26	0.02
logit(Risk score)	−1.47	−1.48	−2.14	0.85	0.01

**Table 3.** Out-of-sample predicted mortality compared against observed mortality in the matched observational study of low and high volume hospitals.

	Low volume	High volume matched	High volume all
Observed mortality	0.2834	0.1982	0.1236
(C,C)	0.2311	0.2158	0.1273
(L,C)	0.2842	0.2069	0.1233
(S,L)	0.2965	0.2081	0.1239
(SLI,L)	0.2961	0.2103	0.1240

reason to be concerned that some important covariate has not been measured, so that a comparison that corrects for measured covariates will not correctly estimate the effect under study. This is a genuine problem in ranking hospitals, and the best solution is to improve the quality of the data used in ranking hospitals. The problem affects both the Bayesian model and the elementary observational comparison that follows in much the same way—neither method addresses potential biases from failure to control an unmeasured covariate. This issue, though both real and important, is less relevant when the focus is on calibration. Calibration asks whether the model's predictions agree with an examination of the data that does not rely upon the model. The model is judged calibrated if its predictions are in reasonable agreement with the elementary observational comparison. The two answers may agree yet both be mistaken estimates of the effects of going to low versus high volume hospitals; that is an important question, but not a question about calibration with the observed data.

Table 2 gives covariate means before and after matching, together with differences in means as a fraction of the standard deviation before matching. Notably, the patients at low and high volume hospitals differed substantially prior to matching, but were similar in matched samples. Patients at low volume hospitals were older on average (84 years vs. 78 years old), with a higher estimated mean probability of death based on patient risk factors (0.22 vs. 0.13), a higher proportion of dementia (22% vs. 12%), a higher proportion with a history of pneumonia (24% vs. 12%), and a somewhat higher history of congestive heart failure (21% vs. 14%), all factors that generally increase mortality risk. Patients at low volume hospitals also had a lower history of prior percutaneous transluminal coronary angiography (prior PTCA) or stenting procedures involving the heart (6% vs. 16%), the history of which generally lowers risk; lower rates of documented atherosclerosis, a cardiac risk factor; and lower rates of anterior infarction, a factor also generally associated with worse prognosis. Part of the difference in mortality between low and high volume hospitals reflects the sicker patient population at low volume hospitals; however, the matching has made an effort to remove this pattern to the extent that it is visible in measured covariates.

The final two columns of Table 2 give standardized measures of covariate imbalance before and after matching. The standardized difference is the difference in means, low volume minus high volume, divided by the standard deviation of the covariate prior to matching. The standard deviation prior to matching is based on the 1302 patients at low volume hospitals and the 50,278 patients at high volume hospitals, pooling the within group variances with equal weights; see Rosenbaum and Rubin (1983) for discussion of this measure of covariate imbalance. For example, the difference in mean ages before matching, 84.3

versus 77.7, is 80% of the standard deviation, but after matching this drops to 1% of the same standard deviation. All of the standardized differences after matching are less than 10% of the standard deviation, whereas many were much larger before matching. In short, the groups look comparable in terms of measured covariates after matching.

If we did not have the Bayes model for individualized prediction, we might rely on a matched observational study to provide general advice. As seen in Table 3, an out-of-sample matched observational study making no use of the Bayes model records a 30-day mortality rate of 28.3% at low volume hospitals, and a mortality rate of 19.8% among similar matched patients at high volume hospitals, or an excess mortality of about 8.5% at low volume hospitals, which is consistent with what the literature says. If one had the option, good general advice would be to seek care for an AMI at a large volume hospital because the mortality rates are lower for patients who look similar in measured covariates describing patients prior to admission.

The remainder of Table 3 sets aside the actual out-of-sample mortality, and instead uses the Bayes models to predict the mortality of the very same patients used in the matched observational study. Let us consider which Bayes models come close to correct predictions, making individual predictions that aggregate to agree with empirically based general advice.

The (C,C) model is very inaccurate in its predictions. That model assumed hospital mortality is independent of volume, and its predictions agree with its assumptions rather than with the out-of-sample data. It says, incorrectly, that mortality is only slightly elevated at low volume hospitals, and it also overstates the mortality at high volume hospitals. In sharp contrast, every one of the other models agree with the general advice that risk is elevated at low volume hospitals. Compared to the (C,C) model, their aggregate predictions are much closer to the actual mortality levels of both the low volume hospitals and their matched counterparts at the high volume hospitals. It is interesting to note that although the overall out-of-sample performance of the (L,C) model was the weakest of the non-(C,C) models in Table 1, its aggregate predictions for the low volume hospitals were better than the rest here. A second illustration of our general out-of-sample calibration approach is presented in Appendix A.3.

The lessons of Table 3 are summarized below.

- It is important to check models against data in a manner that is capable of judging their inadequacies. In the current context, it is difficult to judge that a model is inadequate by predicting the mortality experience of three patients at a hospital whose total AMI volume is three patients. Something else needs to be done to check such predictions.
- It is important to check aspects of models that we actually care about. A spline is an approximation and no one really cares whether it is true or false; rather, we care whether it is

adequate or inadequate as an approximation for something else that we do really care about. In the current context, we care about model predictions that might both affect hospital choice and patient mortality. In particular, a model that says low volume hospitals are safe for AMI treatment when they are not, is a model that is failing in a way that we actually care about.

- A good model for individualized predictions should produce predictions that are, in aggregate, consistent with sound empirically based general advice that we would otherwise fall back on in the absence of individualized predictions. The model should correctly predict the results of sound, out-of-sample studies of general advice that make no use of the model. In Table 3, our models do this, and the (C,C) model does not.
- It is popular to associate Stein's paradox with Bayes inference, but they actually point in different directions. Stein's paradox is a paradox because it seems to say that shrinkage is never harmful providing at least three parameters are estimated; however, it actually refers to a very special situation. In contrast, there is nothing in Bayes inference that suggests one will get the right answer by assuming things that are false or by fitting the wrong model. That the Bayesian, like the classical frequentist, can be wrong, that both need to look at the data to avoid being wrong, to look at the data to judge whether their assumptions are reasonable, and their models are adequate—this need to look at the data—is a strength of the Bayesian and classical frequentist perspectives, not a weakness.

## 5. Standardized Mortality Rates For Public Reporting

After modeling mortality rates as a function of hospital and patient attributes, the next major step in preparing hospital rate estimates for public reporting and further analysis is to remove patient case-mix effects with some form of standardization. Devoid of differences due to patient risk factors, such estimates allow for much clearer assessments of hospital quality. In this vein, Hospital Compare employs a form of indirect standardization for their (C,C) model estimates. As an alternative, we propose a direct standardization approach that more successfully eliminates patient case-mix effects over a wider range of models, and is better calibrated with the actual overall observed mortality rates. Let us proceed to describe and illustrate these two different approaches in detail.

To begin with, both standardization approaches make use of expected mortality rate estimates for any patient at any hospital. If the  $h$ jth patient with covariates  $\mathbf{x}_{hj}$  had been treated at hospital  $h^*$ , under any of our models this rate would be given by

$$p_{h^*}(\mathbf{x}_{hj}) = \text{logit}^{-1}(\alpha_{h^*} + \mathbf{x}'_{hj}\boldsymbol{\beta}), \quad (19)$$

where  $\alpha_{h^*}$  is now the hospital effect and  $\mathbf{x}'_{hj}\boldsymbol{\beta}$  is the usual patient effect. Note that unless  $h^* = h$ , this rate is counterfactual, since patient  $h$ j was actually treated at hospital  $h$ . Note also that for models that include patient-hospital interactions, these interaction covariates would be included as extra columns of  $\mathbf{x}_{hj}$  that change as  $h^*$  is varied. For example, in our (SLIL) model,  $\text{age}_{hj} * \log(\text{vol}_{h^*})$  would be the added interaction covariate

for patient  $h$ j at hospital  $h^*$ . Rather than add cumbersome notation to indicate such dependence of  $\mathbf{x}_{hj}$  on  $h^*$ , for notational simplicity we shall assume that this dependence is implicitly understood from context.

### 5.1. Indirectly Standardized Mortality Rates

As discussed by Ash et al. (2012), as part of its mandate, CMS is charged with quantifying “How does this hospital's mortality for a particular procedure compare to that predicted at the national level for the kinds of patients seen for that procedure or condition at this hospital?” To address this goal, Hospital Compare reports estimates of indirectly standardized hospital mortality rates of the form

$$P_h^{\text{IS}} = (P_h/E_h) \times \bar{y}, \quad (20)$$

where  $E_h$  is an average expected 30-day mortality rate for the hospital  $h$  patients had they been treated at the “national level,” and  $\bar{y}$  is the overall average patient-level mortality rate estimate for AMI. Beyond its intuitive appeal, strictly speaking,  $P_h^{\text{IS}}$  lacks any probabilistic justification as a hospital mortality rate estimate.

For their choice of  $E_h$  in conjunction with the (C,C) model, Hospital Compare uses

$$E_h^{\text{HC}} = \frac{1}{n_h} \sum_j p_{\mu}(\mathbf{x}_{hj}) = \frac{1}{n_h} \sum_{j=1}^{n_h} \text{logit}^{-1}(\mu_{\alpha} + \mathbf{x}'_{hj}\boldsymbol{\beta}), \quad (21)$$

where for patient  $j$  at hospital  $h$ ,  $p_{\mu}(\mathbf{x}_{hj}) = \text{logit}^{-1}(\mu_{\alpha} + \mathbf{x}'_{hj}\boldsymbol{\beta})$  replaces the hospital effect  $\alpha_h$  in  $p_{hj} = \text{logit}^{-1}(\alpha_h + \mathbf{x}'_{hj}\boldsymbol{\beta})$  with the mean hospital effect  $\mu_{\alpha}$  from (2). As opposed to  $p_{hj}$ ,  $\hat{p}_{\mu}(\mathbf{x}_{hj})$  treats every patient as if they went to a hospital with the same average mortality effect  $\mu_{\alpha}$ . To estimate  $E_h^{\text{HC}}$ , Hospital Compare uses SAS's GLIMMIX plug-in estimates of  $\mu_{\alpha}$  and  $\boldsymbol{\beta}$  (Yale New Haven Health Services Corporation 2014, p. 58, eq. (4)).

Although the choice of  $E_h^{\text{HC}}$  for  $E_h$  is reasonable, it is tied directly to the Hospital Compare (C,C) model (2) through  $\mu_{\alpha}$ . To extend indirect standardization beyond the (C,C) model, we propose instead a more flexible and general choice of  $E_h$  that essentially reduces to  $E_h^{\text{HC}}$  under the (C,C) model. Our proposal, which is generally applicable for all the models considered in Section 3, is

$$E_h = \frac{1}{n_h} \sum_{j=1}^{n_h} \left[ \frac{1}{H} \sum_{h^*=1}^H p_{h^*}(\mathbf{x}_{hj}) \right], \quad (22)$$

where from (19),  $p_{h^*}(\mathbf{x}_{hj})$  is the expected mortality rate for the  $h$ jth patient, had they been treated at hospital  $h^*$ . Intuitively,  $E_h$  is the average expected mortality rate of all hospital  $h$  patients had they hypothetically been treated at all hospitals,  $h^* = 1, \dots, H$ . Such averaging over all hospitals removes hospital-to-hospital variation, leaving only the patient attributes to drive the variation of  $E_h$ . With this choice of  $E_h$ , posterior mean Bayes estimates  $\hat{P}_h^{\text{IS}}$  of  $P_h^{\text{IS}}$  in (20) are straightforwardly obtained via the MCMC approach in Section A.1.

Making use of the fact that  $\text{logit}(\cdot)$  is close to linear in the range of most mortality rates here, insight into how  $E_h$  works, as well as its connection with  $E_h^{\text{HC}}$ , is obtained by the

approximation

$$\frac{1}{H} \sum_{h^*} p_{h^*}(\mathbf{x}_{hj}) \approx \text{logit}^{-1}(\bar{\alpha} + \bar{\mathbf{x}}'_{hj} \boldsymbol{\beta}), \quad (23)$$

where  $\bar{\alpha} = \frac{1}{H} \sum_{h^*} \alpha_{h^*}$  and  $\bar{\mathbf{x}}_{hj} = \frac{1}{H} \sum_{h^*} \mathbf{x}_{hj}$ . Recall that  $\mathbf{x}_{hj}$  will vary over  $h^*$  when patient–hospital interaction covariates are present. In models with no patient–hospital interactions, where  $\bar{\mathbf{x}}_{hj} = \mathbf{x}_{hj}$ ,  $E_h$  essentially treats every patient as if they went to a hospital with the same average mortality effect  $\bar{\alpha}$ . In particular, under the (C,C) model where  $\bar{\alpha} \approx \mu_\alpha$  in (21),  $E_h$  will be nearly identical to the Hospital Compare choice  $E_h^{\text{HC}}$ . As a computational shortcut, (23) also provides a convenient route to obtain fast approximations for general  $E_h$ .

To see the effect of the indirect standardization (20) with (22) for the (C,C), (L,C), (S,L), (SLI,I) models, we apply it to obtain the indirectly standardized mortality rate estimates  $\hat{P}_h^{\text{IS}}$  in Figures 4(a)–4(d). In each these plots,  $\hat{P}_h^{\text{IS}}$  has served to transform the mortality rate estimates  $\hat{P}_h$  in Figures 2(a)–2(d) into values that much more closely resemble the hospital effect estimates  $\hat{\alpha}_h$  in Figures 3(a)–3(d). Beginning with the (C,C) model, the plot of the  $\hat{P}_h^{\text{IS}}$  in Figure 4(a) stands in sharp contrast to the plot of the mortality rate estimates  $\hat{P}_h$  in Figure 2(a); note specifically the substantial shrinkage from the two different vertical scales. The effect of dividing  $P_h$  by  $E_h$  has left the  $\hat{P}_h^{\text{IS}}$  estimates looking nearly identical to the hospital effect  $\hat{\alpha}_h$  estimates in Figure 3(a). Evidently, indirect standardization for the (C,C) model has successfully eliminated just about all of the patient case-mix variation from the  $\hat{P}_h$  estimates. Notice also that, unlike the abstract scale of the  $\hat{\alpha}_h$ 's, standardization has left the  $\hat{P}_h^{\text{IS}}$ 's on a mortality scale that makes them easier to interpret and understand.

Turning to the  $\hat{P}_h^{\text{IS}}$  estimates for the (L,C), (S,L), and (SLI,L) models in Figures 4(b)–4(d), we see that these  $\hat{P}_h^{\text{IS}}$ 's have also shrunk the  $\hat{P}_h$ 's to resemble their corresponding  $\hat{\alpha}_h$ 's in Figures 3(b)–3(d). For each model, they are substantially larger at the low volume hospitals, with a clear downward sloping trend. However, unlike the low volume  $\hat{\alpha}_h$ 's in Figures 3(b)–3(d), which are tightly concentrated around their trend averages, the low volume  $\hat{P}_h^{\text{IS}}$  values instead fan out dramatically when  $\text{vol}_h$  is low. Insight into this phenomenon is obtained by considering the ratio of  $P_h$  to the  $E_h$  approximation (23),

$$\frac{P_h}{E_h} \approx \frac{\sum_{j=1}^{n_h} \text{logit}^{-1}(\alpha_h + \mathbf{x}'_{hj} \boldsymbol{\beta})}{\sum_{j=1}^{n_h} \text{logit}^{-1}(\bar{\alpha} + \bar{\mathbf{x}}'_{hj} \boldsymbol{\beta})}. \quad (24)$$

When the  $\alpha_h$  are far from  $\bar{\alpha}$ , as occurs for low volume hospitals under the (L,C), (S,L), and (SLI,L) models, the variation of (24) will reflect the variation of patient attributes  $\mathbf{x}'_{hj}$  across hospitals, especially when  $n_h$  is small. Thus, under these models, the increased variation of the  $\hat{P}_h^{\text{IS}}$  at the low volume hospitals is an artifact of patient-mix variation rather than of the  $\hat{\alpha}_h$  hospital quality variation. Although this phenomenon does not occur under the (C,C) model where all the low volume  $\hat{\alpha}_h$ 's are close to  $\bar{\alpha}$ , this is precisely where the  $\hat{\alpha}_h$ 's estimates were seen to be miscalibrated. For all but the discredited (C,C) model, indirect standardization fails to achieve its goal of eliminating the effect of patient case-mix variation from mortality rates. Furthermore, as we will see in the next section, for every one of our models,

indirectly standardized mortality rates systemically underestimate actual mortality rates. Such indirect standardization cannot be recommended for public reporting.

## 5.2. Directly Standardized Mortality Rates

Directly standardized mortality rates, an alternative to  $P_h^{\text{IS}}$ , directly eliminate patient-mix effects by averaging the mortality rates of all patients had they (hypothetically) been treated at hospital  $h$ . Denoted  $P_h^{\text{DS}}$ , such rates are given by

$$P_h^{\text{DS}} = \frac{1}{N} \sum_{h^*=1}^H \sum_{j=1}^{n_i} p_h(\mathbf{x}_{h^*j}), \quad (25)$$

where  $p_h(\mathbf{x}_{h^*j})$  in (19) is the  $h^*$ th patient's mortality rate had they been treated at hospital  $h$ , and  $N = \sum_{i=1}^H n_i$  is the total number of patients. Because every  $P_h^{\text{DS}}$  is an average over the same set of all patients, there can be no patient-mix differences between them. Note how  $P_h^{\text{DS}}$  is complementary to  $P_h^{\text{IS}}$ , which instead adjusts  $P_h$  by  $E_h$ , the average mortality rate of hospital  $h$  patients had they (hypothetically) been treated at all hospitals. Posterior mean Bayes estimates  $\hat{P}_h^{\text{DS}}$  of  $P_h^{\text{DS}}$  in (25) are straightforwardly obtained via the MCMC approach in Section A.1.

Further insight into  $P_h^{\text{DS}}$  is obtained by the approximation

$$P_h^{\text{DS}} \approx \text{logit}^{-1}(\alpha_h + \bar{\mathbf{x}}'_h \boldsymbol{\beta}), \quad (26)$$

where  $\bar{\mathbf{x}}_h = \frac{1}{M} \sum_{h^*} \sum_j \mathbf{x}_{h^*j}$ , which follows from the fact that  $\text{logit}(\cdot)$  is close to linear in the range of most mortality rates here. Thus,  $P_h^{\text{DS}}$  may be regarded as the expected mortality rate at hospital  $h$  of a patient with average  $\bar{\mathbf{x}}_h$  attributes. In models with no patient–hospital interactions (where  $\mathbf{x}_{h^*j}$  does not vary with  $h$ ),  $\bar{\mathbf{x}}_h$  is simply the mean of  $\mathbf{x}_{h^*j}$  over all patients.

Figures 5(a)–5(d) plots directly standardized mortality rate  $\hat{P}_h^{\text{DS}}$  estimates by  $\text{vol}_h$  for each of the (C,C), (L,C), (S,L), and (SLI,L) models. As with  $\hat{P}_h^{\text{IS}}$ ,  $\hat{P}_h^{\text{DS}}$  has served to transform the  $\hat{P}_h$  in Figures 2(a)–2(d) into values that much more closely resemble the hospital effect estimates  $\hat{\alpha}_h$  in Figures 3(a)–3(d), but that remain on a more meaningful mortality scale. Up to this rescaling, the  $\hat{P}_h^{\text{DS}}$  estimates under the (C,C) model are also virtually identical to their  $\hat{\alpha}_h$ 's, and under the (L,C) and (S,L) models now appear much more similar to their  $\hat{\alpha}_h$ 's. No longer fanning out at the low volume hospitals as the  $\hat{P}_h^{\text{IS}}$  rates did, these  $\hat{P}_h^{\text{DS}}$ 's have more successfully eliminated patient case-mix variability. Indeed, with linear correlations of 0.9967, 0.9978, 0.9975 under these three models,  $\hat{P}_h^{\text{DS}}$  serves as meaningfully interpretable, nearly linear rescaling of the  $\hat{\alpha}_h$  hospital effect estimates. For the (SLI,L) model, the  $\hat{P}_h^{\text{DS}}$ 's are more dispersed than the  $\hat{\alpha}_h$ 's, tracking them less closely with a correlation of 0.9906. This is not surprising because when patient–hospital interactions are present, as we saw in Section 3.4, the  $\hat{\alpha}_h$ 's no longer entirely capture hospital effects. Evidently, the  $\hat{P}_h^{\text{DS}}$ 's are a much more effective reflection of actual hospital effects, and one that puts them on a natural mortality scale.

It is concerning to see that the overall level of the  $\hat{P}_h^{\text{IS}}$  rates in Figures 4(a)–4(d) is systematically lower than the overall level of the  $\hat{P}_h^{\text{DS}}$  rates in Figures 5(a)–5(d). To understand what is going on, we have put two horizontal lines on each of the  $\hat{P}_h^{\text{DS}}$  plots. For each model, the higher line is the simple average of the  $\hat{P}_h^{\text{DS}}$  rates,



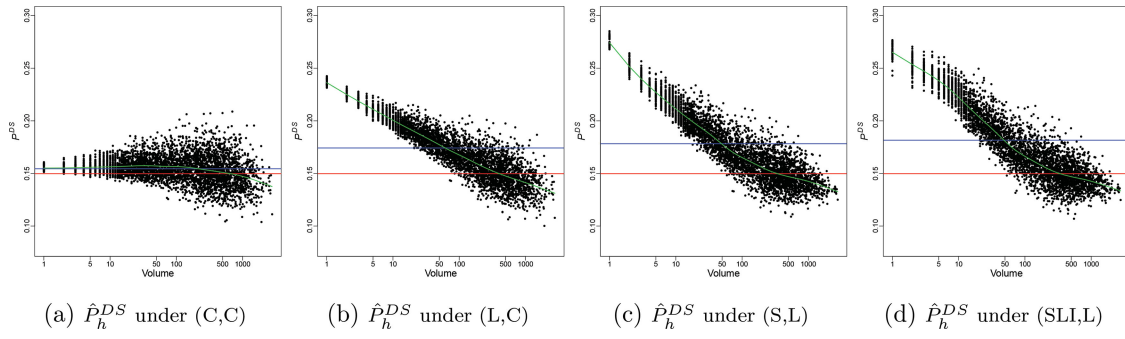


Figure 5.  $\hat{P}_h^{DS}$  versus  $\text{vol}_h$ .

while the lower line is the same observed average mortality rate  $\bar{y} = 0.1498$  obtained by averaging over all patients in our data. By the indirect standardization construction, the simple average of the  $\hat{P}_h^{IS}$  rates will always equal the average patient mortality rate  $\bar{y}$ , as is evident in their plots. In fact, the  $\hat{P}_h^{IS}$  rates understate the poor performance of the worst hospitals by overstating the risk faced by the typical patient as the following discussion will show.

As we have seen, low volume hospitals have higher than typical risk, but treat relatively few patients; therefore, the (unweighted) average risk over hospitals is much higher than the average risk faced by patients. Saying the same thing differently, a random patient likely went to a larger volume hospital—that is, what it means to be a larger volume hospital—but a random hospital is unlikely to have very high volume—that is what it means to be a random hospital. The expected risk  $E_h$  used by  $\hat{P}_h^{IS}$  in approximation (23) is essentially obtained by substituting the average hospital effect  $\bar{\alpha}$  for the specific hospital effect  $\alpha_h$  in the various expressions for patient mortality rates. This average hospital effect  $\bar{\alpha}$  describes the typical hospital, not the hospital that treats the typical patient. Therefore,  $E_h$  is too high: it describes the risk that would be relevant if patients picked hospitals at random with equal probabilities, but they do not; rather, they tend to go to larger volume hospitals with lower risk. This one problem with  $\hat{P}_h^{IS}$  could be fixed with a patch: instead of using the unweighted  $\bar{\alpha}$ , one could average over hospitals with weights proportional to their volumes that would describe the risk faced by the typical patient. However, this would still not resolve the patient-mix variability shortcomings of  $\hat{P}_h^{IS}$  discussed in the previous section. Overall, we recommend  $\hat{P}_h^{DS}$  over  $\hat{P}_h^{IS}$  as a more reliable standardized mortality rate report, especially for the model elaborations that we have proposed.

## 6. Learning from Directly Standardized Mortality Rates

### 6.1. Mortality Rate Uncertainty

For each hospital  $h$ , each of the Bayesian models induces a posterior distribution  $\pi(P_h^{DS} | \mathbf{y})$  on its directly standardized mortality rate. Averages of MCMC simulated samples from each of these posteriors produced the posterior mean estimates  $\hat{P}_h^{DS}$  plotted in Figures 5(a)–5(d). As is strikingly evident in every plot, the hospital-to-hospital variation of these  $\hat{P}_h^{DS}$  values is smallest at the low volume hospitals, gradually increasing as volume increases. This is a consequence of the stronger shrinkage of the  $\hat{\alpha}_h$  estimates to their means by each of the random effects models.

However, this observed hospital-to-hospital variation of the  $\hat{P}_h^{DS}$  values should not be confused with the posterior uncertainty of the accuracy of each estimate, which can be much larger. This uncertainty is captured by the full posterior distribution of each  $\hat{P}_h^{DS}$  value, and can be conveyed with interval estimates based on the simulated posterior samples for each hospital. For example, 95% intervals may be obtained from the 2.5% and 97.5% sample quantiles. Such intervals can be used to provide a direct assessment of the reliability of each  $\hat{P}_h^{DS}$  estimate, a more informative alternative to the practice of eliminating estimates because of small sample sizes or other reliability adjustments, as advocated, for example, by Dimick, Staiger, and Birkmeyer (2010).

To illustrate this uncertainty, Figures 6(a)–6(d) display boxplots of simulated posterior samples from the  $\pi(P_h^{DS} | \mathbf{y})$  distributions for 10 typical hospitals of sizes  $\text{vol}_h = 1, 2, 5, 10, 25, 50, 100, 200, 400, 800$  under the (C,C), (L,C), (S,L), and (SLI,L) models, respectively. Notice how the profile

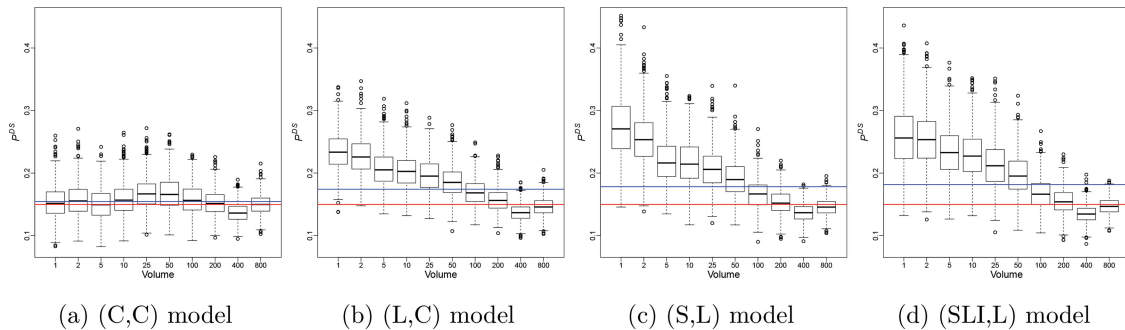


Figure 6.  $\hat{P}_h^{DS}$  posterior uncertainty at 10 hospitals of varying volume.

**Table 4.** Hospital classifications by mortality rates.

Counts (%)	All hospitals				Lower volume quartile hospitals				Upper volume quartile hospitals			
	Low	Average	High	Total	Low	Average	High	Total	Low	Average	High	Total
(C,C)	33 (0.752)	4333 (98.57)	30 (0.68)	4396	0 (0.00)	1116 (100.00)	0 (0.00)	1116	32 (2.91)	1047 (95.27)	20 (1.82)	1099
(SLI,L)	58 (1.32)	3310 (75.30)	1028 (23.28)	4396	0 (0.00)	210 (18.82)	906 (81.18)	1116	57 (5.19)	1038 (94.45)	4 (.036)	1099

of mortality rate uncertainty under the (C,C) model stands out from the rest. Under the (C,C) model, mortality rate uncertainty is hardly related to volume, the level, and spread of the posterior distributions being roughly the same at the smaller  $\text{vol}_h$  values. But under our models, especially the fully emancipated (S,L) and (SLI,L) models, both the level and spread of the posterior distributions are higher for the low volume hospitals, decreasing steadily as  $\text{vol}_h$  increases.

To further illustrate the informative value of mortality rate posterior reporting under our approach, Silber et al. (2016) applied a hospital attribute enhanced variant of our (L,C) model to estimate mortality rates at five hospitals in Chicago, IL. Consistent with Figure 6, mortality rate posteriors for the smaller volume hospitals there are seen to be both higher and more diffuse under the enhanced (L,C) model than under the (C,C) model.

Thus, the systematically higher mortality rate estimates at low volume hospitals under our models are also each less precise or reliable in the sense that there is more uncertainty about their accuracy. Figure 6 reminds us that in judging the mortality rate estimate for a given hospital with our models, consideration must be given both to the point estimate and its uncertainty as reflected by the posterior distribution. In particular, if a small hospital was plausibly excellent, an analysis of this form would convey that such excellence is plausible.

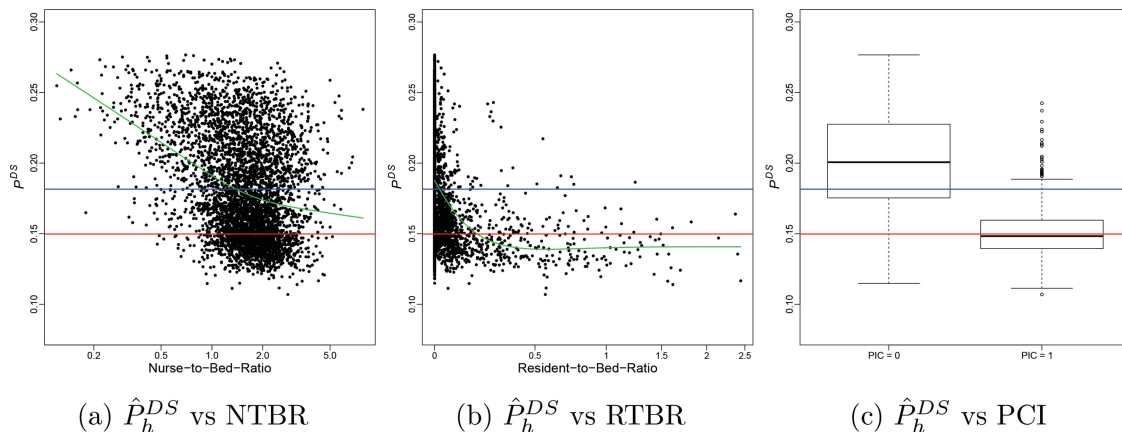
## 6.2. Hospital Classification by Mortality Rates

Do many or few hospitals have high mortality rates compared to national rates? The 95% credibility intervals for  $\hat{p}_h^{DS}$  described in the previous section can be used to classify hospitals as low, average, or high mortality according to whether its 95% interval is entirely below, intersects, or is entirely above the overall average mortality rate of 15%.

Table 4 provides these classifications for the (C,C) and the (SLI,L) models. Overall, the (C,C) model categorizes most hospitals, 4333, as average mortality, with only 33 as low and 30 as high. Our (SLI,L) model is much more discriminating, classifying 3310 hospitals as average, with 58 as low and 1028 as high. However, much of this discrimination between average and high mortality hospitals occurs at the lower volume quartile hospitals. Whereas all 1116 low volume quartile hospitals are classified as average mortality by the (C,C) model, the (SLI,L) has recategorized 906 of these as high mortality. For the higher volume quartile hospitals, the (C,C) model classifies 32 as low, 20 as high, whereas the (SLI,L) model has many more, 57 as low, and only 4 as high. The full cross-classification leading to Table 4 appears as Table A.1 in Appendix A.4. There we see, for example, that of the 57 higher volume hospitals classified as low mortality by (SLI,L), 28 were classified as average by (C,C). In contrast, of the 32 higher volume hospitals classified as low mortality by (C,C), only 4 were classified as average by (SLI,L).

## 6.3. The Influence of Hospital Attributes

The (SLI,L) model included three attributes of hospitals besides volume, namely, NTBR, RTBR, and PCI. The 95% highest posterior density intervals for the coefficients of NTBR and RTBR included only negative values and excluded zero, while the interval for PCI included zero. However, PCI is highly correlated with hospital volume, which is also in the model in the form of a spline. Figures 7(a)–7(c) plots  $\hat{p}_h^{DS}$  versus NTBR, RTBR, and PCI, respectively, for this model. In Figures 7(a) and 7(b), hospitals with more nurses per bed or more residents per bed are predicted to have lower mortality. Although not confirmed to be distinct from volume, the ability to perform PCI (PTCA, stents, or CABG) is also associated with lower mortality in Figure 7(c). These patterns are generally consistent with the health services

**Figure 7.**  $\hat{p}_h^{DS}$  under the (SLI,L) model.

research literature concerning the influence of invasive cardiology (Stukel et al. 2007) and nurse staffing (Person et al. 2004) on AMI survival, though the benefit of treatment at a teaching hospital is more controversial, and studies are inconsistent regarding its influence on mortality (Allison et al. 2000; Navathe et al. 2013). Lastly, for the age–volume interaction that was also included in the (SLI,L) model, a positive coefficient posterior estimate indicated that large volume hospitals confer a greater survival benefit for younger medicare patients.

## 7. Discussion

As a model for AMI hospital mortality rates, we have found the hierarchical random effect logit model used by Hospital Compare to be inadequate compared to alternatives that model hospital effects as a functions of hospital attributes. Such models were seen to offer substantial predictive improvements as measured by out-of-sample predictive Bayes factors. Going further, we have suggested calibrating individualized predictions from Bayesian models against empirically based general advice that would otherwise be used to inform decisions. This entails conducting an out-of-sample study of general advice without using the model, and then using the model to predict the results of that independent study. For this purpose, a matched out-of-sample comparison confirmed familiar advice that low volume hospitals tend to have higher mortality rates when treating AMI. While our models accurately predicted the results of that matched out-of-sample study, the current Hospital Compare model is not calibrated in this sense, and hence should not be used.

One of the main goals of this article has been to show that the inclusion of hospital attributes in our models leads to better calibrated and more informative mortality rate predictions. In particular, we obtained a vast improvement over the Hospital Compare model with the Medicare data by including hospital volume, staffing by nurses and residents, PCI, and patient-age by hospital-volume interactions. While this improvement better serves the needs of patients, its ramifications for public policy must be considered cautiously. Because low volume hospitals, by definition, have little data regarding mortality, a better small volume hospital may be unable to overcome the poor results of its similarly sized peers to receive the good ranking it deserves. To some extent, this problem can be mitigated with our models, by including measures of uncertainty along with mortality rate estimates, as described in Section 6.1. However, further modeling with additional hospital attributes has the clear potential to shed more light on the rankings of such hospitals. In particular, the addition of hospital attributes that distinguish better low volume hospitals from the rest would be ideal. Indeed, our models should be considered as the beginning rather than the end of the story. While hospital volume is a convenient variable, which is strongly associated with mortality, with more information, for example, about hospital management, it may even turn out not to be the most important predictive variable, thereby diminishing its penalizing effect.

We have also recommended that the common practice of reporting indirectly standardized rates be avoided. We found that for our improved models, indirectly standardized rates fail to eliminate the effect of patient mix differences across hospitals. Furthermore, such indirectly adjusted rates are also

inherently misleading in that they systematically underestimate population hospital mortality rates. In contrast, direct adjustment faithfully translates the model's adjustments and mortality predictions into a properly calibrated and easily understood format for public reporting.

The future lies with predictions that are individualized not just to particular hospitals, but to particular patients when treated at particular hospitals. We scratched the surface of this topic by including patient-by-hospital interactions in (19), finding that younger Medicare patients benefit more than older Medicare patients from treatment at high volume hospitals. Note that once we have fit one of our models and obtained estimates of the  $\alpha_h$  and  $\beta$ , (19) can be applied with  $p_h(\mathbf{x})$  for any patient characteristics  $\mathbf{x}$  to obtain mortality rate estimates for that patient at any hospital. Such personalized rate estimates would be more relevant than standardized rates for any particular patient.

## Supplementary Materials

Appendix A.1 details the MCMC implementation for simulated sampling from our hierarchical logit model posteriors. This entails successive substitution Gibbs sampling from the full conditionals obtained with a suitable Poly-Gamma latent variable posterior augmentation. Appendix A.2 illustrates the relationship between mortality rates and number of hospital beds with a model that excludes volume. This gives further insight into the persistent relationship between hospital mortality and hospital size. Appendix A.3 provides a second out-of-sample calibration example with US News and World Report hospital rankings. Appendix A.4 provides the full cross-classification of low, average and high hospital mortality rates by the (C,C) and the (SLI,L) models.

## Acknowledgments

The authors are especially grateful to Nabanita Mukherjee, an associate editor, and anonymous referees for their many constructive suggestions.

## Funding

This work was supported by the Agency for Healthcare Research and Quality grant No. R21-HS021854; and grants SBS-1260782 and DMS-1406563 from the National Science Foundation.

## References

- Allison, J. J., Kiefe, C. I., Weissman, N. W., Person, S. D., Rousculp, M., Canto, J. G., Bae, S., Williams, O. D., Farmer, R., and Centor, R. M. (2000), "Relationship of Hospital Teaching Status with Quality of Care and Mortality for Medicare Patients with Acute MI," *Journal of the American Medical Association*, 284, 1256–1262. [946]
- Ash, A., Fienberg, S. E., Louis, T. L., Normand, S. T., Stukel, T. A., and Utts, J. (2012), "Statistical Issues in Hospital Performance," Commissioned by the Committee of Presidents of Statistical Societies. Available at <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf>. [942]
- Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer Verlag. [936]
- (2006), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1, 385–402. [939]
- Box, G. E. P., and Meyer, R. D. (1986), "Dispersion Effects from Fractional Designs," *Technometrics*, 28, 19–27. [938]
- Cox, D. R. (1961), "Tests of Separate Families of Hypotheses," in *Proceedings of the Fourth Berkeley Symposium*, pp. 105–123. [939]

- Dawid, A. P. (1982), "The Well-Calibrated Bayesian," *Journal of the American Statistical Association*, 77, 605–610. [934]
- Dimick, J. B., Staiger, D. O., and Birkmeyer, J. D. (2010), "Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment," *Health Services Research*, 45, 1614–1629. [944]
- Gandjour, A., Bannenberg, A., and Lauterbach, K. W. (2003), "Threshold Volumes Associated with Higher Survival in Health Care: A Systematic Review," *Medical Care*, 41, 1129–1141. [934,936]
- Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society, Series B*, 56, 501–514. [939]
- Grieco, N., Ieva, F., and Paganoni, A. M. (2012), "Performance Assessment Using Mixed Effects Models: A Case Study on Coronary Patient Care," *IMA Journal of Management Mathematics*, 23, 117–131. [939]
- Gu, Y., Fiebig, D. G., Cripps, E., and Kohn, R. (2009), "Bayesian Estimation of A Random Effects Heteroscedastic Probit Model," *Econometrics Journal*, 12, 324–339. [938]
- Guglielmi, A., Ieva, F., Paganoni, A. M., Ruggeri, F., and Soriano, J. (2014), "Semiparametric Bayesian Modeling for the Classification of Patients with High Observed Survival Probabilities," *Journal of the Royal Statistical Society, Series C*, 63, 25–46. [939]
- Halm, E. A., Lee, C., and Chassin, M. R. (2002), "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature," *Annals of Internal Medicine*, 137, 511–520. [934,936]
- Hansen, B. B. (2007), "Optmatch: Flexible, Optimal Matching for Observational Studies," *R News*, 7, 18–24. [940]
- (2008), "Prognostic Analogue of the Propensity Score," *Biometrika*, 95, 481–488. [940]
- Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [939]
- Krumholz, H. M., Wang, Y., Mattera, J. A., Wang, Y., Han, L. F., Ingber, M. J., Roman, S., and Normand, S.-L. T. (2006), "An Administrative Claims Model Suitable for Profiling Hospital Performance Based on 30-Day Mortality Rates among Patients with an Acute Myocardial Infarction," *Circulation*, 113, 1683–1692. [935]
- Luft, H. S., Hunt, S. S., and Maerki, S. C. (1987), "The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns?" *Health Services Research*, 22, 157–182. [934,936]
- Navathe, A. S., Silber, J. H., Zhu, J., and Volpp, K. G. (2013), "Does Admission to a Teaching Hospital Affect Acute Myocardial Infarction Survival?" *Academic Medicine*, 88, 475–482. [946]
- Person, S. D., Allison, J. J., Kiefe, C. I., Weaver, M. T., Williams, O. D., Centor, R. M., and Weissman, N. W. (2004), "Nurse Staffing and Mortality for Medicare Patients with Acute Myocardial Infarction," *Medical Care*, 42, 4–12. [946]
- Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349. [935]
- Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer. [940]
- Rosenbaum, P. R., and Rubin, D. B. (1983), "Constructing a Control Group by Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician*, 39, 33–38. [941]
- Shahian, D. M., and Normand, S. L. (2003), "The Volume-Outcome Relationship: From Luft to Leapfrog," *Annals of Thoracic Surgery*, 75, 1048–1058. [934,936]
- Silber, J. H., Rosenbaum, P. R., Brachet, T. J., Ross, R. N., Bressler, L. J., Even-Shoshan, O., Lorch, S. A., and Volpp, K. G. (2010), "The Hospital Compare Mortality Model and the Volume-Outcome Relationship," *Health Services Research*, 45, 1148–1167. [934]
- Silber, J. H., Satopää, V. A., Mukherjee, N., Ročková, V., Wang, W., Hill, A. S., Even-Shoshan, O., Rosenbaum, P. R., and George, E. I. (2016), "Improving Medicare's Hospital Compare Mortality Model," *Health Services Research*, 51, 1229–1247. [945]
- Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., and Grigg, O. (2012), "Statistical Methods for Healthcare Regulation: Rating, Screening and Surveillance," *Journal of the Royal Statistical Society, Series A*, 175, 1–47. [939]
- Stuart, E. A. (2010), "Matching Methods for Causal Inference," *Statistical Science*, 25, 1–21. [940]
- Stukel, T. A., Fisher, E. S., Wennberg, D. E., Alter, D. A., Gottlieb, D. J., and Vermeulen, M. J. (2007), "Analysis of Observational Studies in the Presence of Treatment Selection Bias: Effects of Invasive Cardiac Management on AMI Survival Using Propensity Score and Instrumental Variable Methods," *Journal of the American Medical Association*, 297, 278–285. [946]
- Yale New Haven Health Services Corporation. (2014), *Measures Updates and Specifications (Version 8.0)*, Washington, DC: Centers for Medicare and Medicaid Services. [935,942]