STATISTICAL ISSUES IN ASSESSING HOSPITAL PERFORMANCE

Commissioned by the Committee of Presidents of Statistical Societies

The COPSS-CMS White Paper Committee:

Arlene S. Ash, PhD; Stephen E. Fienberg, PhD; Thomas A. Louis, PhD
Sharon-Lise T. Normand, PhD; Thérèse A. Stukel, PhD; Jessica Utts, PhD

Original report submitted to CMS on November 28, 2011
Revised on January 27, 2012

Preface

The Centers for Medicare and Medicaid Services (CMS), through a subcontract with Yale New Haven Health Services Corporation, Center for Outcomes Research and Evaluation (YNHHSC/CORE), is supporting a committee appointed by the Committee of Presidents of Statistical Societies (COPSS) to address statistical issues identified by the CMS and stakeholders about CMS's approach to modeling hospital quality based on outcomes. In the spring of 2011, with the direct support of YNHHSC/CORE, COPSS formed a committee comprised of one member from each of its constituent societies, a chair, and a staff member from the American Statistical Association, and held a preliminary meeting in April. In June, YNHHSC/CORE executed a subcontract with COPSS under its CMS contract to support the development of a White Paper on statistical modeling. Specifically, YNHHSC/CORE contracted with COPSS to "provide guidance on statistical approaches ... when estimating performance metrics," and "consider and discuss concerns commonly raised by stakeholders (hospitals, consumer, and insurers) about the use of "hierarchical generalized linear models in profiling hospital quality. The committee convened in June and August of 2011, and exchanged a wide variety of materials. To ensure the committee's independence, YNHHSC/CORE did not comment on the white paper findings, and CMS pre-cleared COPSS' publication of an academic manuscript based on the White Paper.

The committee thanks COPSS and especially its chair, Xihong Lin of the Harvard School of Public Health; and staff of the American Statistical Association, especially Steve Pierson and Keith Crank, for their efforts in establishing the committee and coordinating its work. We thank Darcey Cobbs-Lomax and Elizabeth Drye of the Yale Center for Outcomes Research and Evaluation (CORE), Yale New Haven Hospital who issued the contract on behalf of CMS.

COPSS developed a special formal review process for this report with the goals of ensuring that it is objective and addresses the CMS charge. Consequently, this report was reviewed in draft form by professionals with a broad range of perspectives and expertise. Xihong Lin coordinated the review. We thank her and the following individuals for donating their time and expertise: Adalsteinn Brown, University of Toronto; Jim Burgess, Boston University; Justin Dimick, University of Michigan; Frank Harrell, Vanderbilt University; Jack Kalbfleisch, University of Michigan; Catarina Kiefe, University of Massachusetts; Niek Klazinga, University of Amsterdam; Neil Prime, Care Quality Commission, UK; Susan Paddock, The RAND Cooperation; Patrick Romano, University of California at Davis; David Spiegelhalter, University of Cambridge; Robert Wolfe, University of Michigan; Alan Zaslavsky, Harvard Medical School; and three anonymous reviewers.

The Committee of Presidents of Statistical Societies (COPSS)

Charter member societies of the Committee of Presidents of Statistical Societies (COPSS) are: the American Statistical Association, the Eastern North American Region of the International Biometric Society, the Institute of Mathematical Statistics, the Statistical Society of Canada, and the Western North American Region of the International Biometric Society. COPSS leadership consists of the Chair, Secretary/Treasurer, Presidents, Past Presidents, and Presidents-Elect of the charter member societies.

The preamble to the COPSS charter states,

> "Whereas the various societies have distinct characteristics they also have some common interests and concerns that can benefit from coordinated effort. The purpose of the Committee of Presidents of Statistical Societies (COPSS) is to work on shared problems, to improve intersociety communication, and to offer distinguished awards. Other activities designed to promote common interests among the member societies may be undertaken from time to time."

See, (`http://nisla05.niss.org/copss/`) for additional information.

COPSS-CMS White Paper Committee Members

(Appendix B contains biographical information)

ARLENE S. ASH, PHD; Professor, Department of Quantitative Health Sciences; University of Massachusetts Medical School; Worcester, MA USA
(Representing the *American Statistical Association*)

STEPHEN E. FIENBERG, PHD; Maurice Falk University Professor of Statistics and Social Science; Department of Statistics; Carnegie Mellon University; Pittsburgh, PA USA
(Representing the *Institute of Mathematical Statistics*)

THOMAS A. LOUIS, PHD; Professor, Department of Biostatistics; Johns Hopkins Bloomberg School of Public Health; Baltimore, MD USA
(*Committee Chair*)

SHARON-LISE T. NORMAND, PHD; Professor, Department of Health Care Policy, Harvard Medical School; and Department of Biostatistics, Harvard School of Public Health; Boston, MA USA
(Representing the *Eastern North American Region of the International Biometric Society*)

THÉRÈSE A. STUKEL, PHD; Senior Scientist, Institute for Clinical Evaluative Sciences; Professor, Institute of Health Policy, Management & Evaluation; University of Toronto, Canada; Professor, Dartmouth Institute for Health Policy and Clinical Practice; Hanover, NH, USA
(Representing the *Statistical Society of Canada*)

JESSICA UTTS, PHD; Professor of Statistics; University of California, Irvine; Irvine, CA USA
(Representing the *Western North American Region of the International Biometric Society*)

Contents

Executive Summary

The Centers for Medicare and Medicaid Services (CMS) charged the committee to:

> "Provide guidance on statistical approaches for accounting for clustering and variable sample sizes across hospitals when estimating hospital-specific performance metrics (e.g., mortality or readmission rates)."

and to

> "Consider and discuss concerns commonly raised by stakeholders (hospitals, consumers, and insurers) about the use of HGLMs [Hierarchical Generalized Linear Models] in public reporting of hospital quality."

The committee addresses this charge and related issues with the goal of enhancing the validity, credibility, and clarity of the CMS evaluations of hospital performance. In doing so the committee began by interpreting CMS's broad goal to be:

> Provide hospital-specific performance metrics for an array of procedures that incorporate the best possible information for each hospital as to how well it performs with its patients in comparison to the outcomes that would be expected if the same patients were to receive care that matched the national norm.

Given CMS's congressional mandate, these metrics clearly involve point estimates of performance, in the form of a standardized mortality rate, and assessment of the uncertainty associated with such estimates.

The committee reviews the hierarchical modeling approach to performance measures based on the concept of a standardized mortality ratio, and contrasts the current CMS approach with other approaches proposed in the literature. The report describes the assumptions underlying different methods and the extent to which there is empirical support for them. The report sets this discussion in the broader context of statistical methods for a variety of other purposes, especially in the context of large sparse data sets, and includes suggestions for improving upon the current CMS method.

The following Commentary and Recommendations duplicate section 11 of the full report.

Commentary on principal criticisms of the current CMS approach

The committee has addressed the criticisms received by the CMS in response to the use of hierarchical logistic regression modeling in measure development as follows:

Criticism 1: The approach fails to reveal provider performance variation: The hierarchical modeling shrinkage effect reduces reported variation of hospital performance and renders the information not useful for consumers.

> *Committee view:* The CMS seeks to report on systematic differences in patient outcome due to hospital quality, after removing variability in observed outcomes that is due to differences in case mix and stabilizing highly variable estimates. Even after risk adjustment for case-mix differences, inherent randomness causes directly estimated hospital effects and relative

rates (that is, O/E ratios where the observed rate (O) is divided by its national model based expected rate (E)) for some hospitals to vary more than the systematic effects that are to be identified. This is especially true for hospitals with extremely low volumes, whose ratios provide little information about their underlying relative rates due to having very wide confidence intervals. Large reductions in reported variation are appropriate for hospital performance measures where the true systematic differences across all hospitals are small. The committee identifies as a top priority evaluating the option of expanding the model to include shrinkage targets that depend on hospital attributes.

Criticism 2: The approach masks performance of small hospitals: It is pointless to include small (low volume) hospitals in the calculations based on hierarchical modeling because they would get a rate close to the national mean. The hierarchical modeling methodology neutralizes small hospital performance.

Committee view: Data from small hospitals provide considerable information on associations between patient case mix and the outcome for parameter estimation in the hierarchical model; therefore, their data should be included in model building. The standard errors of hospital-specific estimates for low volume hospitals are typically large. Stabilization requires that these highly variable estimates are moved towards a model-based target to a greater degree than less variable estimates, resulting in more shrinkage for low volume hospital estimates. The overarching goal is to produce estimates that better reflect true, underlying hospital effects. As stated in the response to criticism 1, the committee identifies as a top priority evaluating the option of expanding the model to include shrinkage targets that depend on hospital attributes.

Criticism 3: The approach is based on complicated concepts and is difficult to communicate and explain to the public and to the providers. Stakeholders are familiar with the numerator and the denominator (O/E) and the output of logistic regression modeling, but the approach adopted by CMS replaces the "O" with a shrinkage estimate (referred to as "predicted" in CMS documents), and the concept is difficult to convey. In addition the concept of a hospital-specific effect is not comprehensible to most of the stakeholders.

Committee view: Some concepts and computations are more complicated than for the standard logistic regression approach, but the additional complexity allows for respecting the hierarchical structure of the data and stabilizing estimates, thereby reducing regression to the mean effects and bouncing around of provider-specific estimates. Furthermore, a fixed-effects, logistic regression model also produces hospital-specific effects. There is a continuum between the single-intercept, random effects model and the fixed-effects model with a directly estimated intercept for each hospital, with the middle ground being occupied by a mixed effects model that includes hospital-level covariates. Therefore, barriers to comprehension of the concept are shared by all approaches. This report clarifies the principal building blocks of the approaches. The committee calls for improved communication on goals, methods, and interpretations.

Criticism 4: The evaluation of the National Quality Forum (NQF) steering committees on use of hierarchical modeling has been inconsistent, contingent on the point of view of the panelists. Therefore, it shows a lacks of consensus among statisticians and health service researchers in using hierarchical modeling for risk adjustment of outcome measures.

Committee view: The committee notes that to make progress on this issue, and possibly come to consensus, the debate must be evidence-based starting with a clear articulation of goals and ending with effective evaluation of the properties of candidate approaches. The committee recommends use of hierarchical models as an effective method to account for clustering of admissions within providers, to support valid and effective risk adjustment, and to produce stabilized estimates, although it recognizes that other approaches can accomplish these goals. This report clarifies goals and why hierarchical models are a valid approach, and focuses discussion on potential enhancements of the current CMS method.

The committee's investigation has led to the following conclusions and recommendations:

1. *Use of Hierarchical Generalized Linear Models*

   The committee concludes that Hierarchical Generalized Linear Logistic Modeling is an effective analytic approach that accounts for the structure of the data used in CMS mortality and readmission hospital metrics. The approach accommodates modeling of the association between outcomes and patient-level, pre-admission characteristics; with appropriate inclusion of hospital-level attributes, it can adjust the patient-outcome relation for potential confounding by hospital to the degree that the necessary information is available; and supports stabilizing hospital-specific performance estimates by shrinking direct estimates towards an appropriate target. The amount of shrinkage can be controlled to the extent that these controls accord with CMS' primary goal (see Recommendations 3, 4, and 5 below).

2. *Incorporation of procedure-specific volume*

   Other recommendations encourage serious consideration of including hospital-level (not procedure-specific) attributes in the national-level, case mix adjustment model and in setting shrinkage targets for stabilizing estimated hospital effects. The committee cautions that the issues related to use of procedure-specific volume are complex. Volume has a combined role as both an exogenous attribute that may be an independent predictor of quality (e.g., practice makes perfect) and an endogenous attribute that is in the causal pathway of the outcome. Furthermore, "low procedure-specific volume" may be a marker for an inadequate risk adjustment that disadvantages hospitals with low-volume procedures.

   Though evaluation of including procedure-specific volume is important, the committee recommends that higher priority be given to use of other hospital-level attributes in modeling case-mix and in producing shrinkage targets. However, regarding volume, use

of procedure-specific volume from time periods prior to those used in an assessment is likely not problematic and should be explored for its ability to contribute to better-tailored shrinkage targets.

3. *Case-mix adjustment*

   (a) Patient-level attributes

      i. Consider whether the current set of patient-level attributes should be augmented, for example by including race or other demographics.

      ii. Evaluate broadening modeling approaches to include additional interaction terms among patient-level attributes.

      iii. Evaluate further broadening patient-level models through use of splines, classification and regression trees, random forests, and boosting (see section 8) to see if relative to current approaches they improve case mix adjustments by producing predictions with lower mean squared error, or improve other performance measures such as those in Efron (1978).

   It will be important to explore the extent to which alternative modeling strategies improve case-mix adjustments by producing predictions with lower mean squared error (i.e., predictions that are closer to the true structural relation), or improve other statistical attributes.

   (b) Hospital-level attributes

      The committee recommends that the CMS explore how best to include hospital attributes for two distinct purposes: 1) when developing the national-level risk model to reduce potential confounding induced by correlation between hospital and patient-level attributes; and 2) when calculating the shrinkage targets used to stabilize SMRs. Incorporating them is an accepted approach in other facility assessment settings. It is very important to note that hospital-level attributes should not set the comparator for a hospital's performance; indeed, the denominator of the SMR should depend only on a validly estimated relation between patient-level attributes and outcome. However, there may be confounding of this relationship with certain hospital characteristics, and methods to reduce this confounding should be explored.

      To reduce confounding and stabilize hospital-specific estimates, the committee proposes in appendix F.1 a Bayesian hierarchical model that adjusts for hospital-level attributes when developing the risk model, but constrains risk predictions to be for a "typical hospital" so that hospital-level attributes play no other role in producing the expected value for a hospital. The model also allows for hospital-attribute-specific shrinkage targets to stabilize estimated SMRs.

      The committee cautions that although statistical models are available to accomplish these goals, decisions as to what attributes to include and how to include them must

be carefully considered. For example, covariate interactions may be needed (e.g., between hospital size and rural/urban status). Coding for candidate attributes needs to be evaluated. For example, should size be retained as a continuous attribute or be categorized? If categorized, how should the number of categories, and their cutoff values, be determined?

4. *Stabilizing estimated hospital effects*

   (a) Including hospital-level attributes in determining the shrinkage target when stabilizing estimated hospital effects is standard practice in other facility assessment settings. The committee recommends that the CMS give serious consideration to including such variables in setting shrinkage targets. To reduce potential confounding, covariate main effects and possibly interactions should be considered (e.g., shrinkage targets could be different for each of small-rural, large-rural, small-urban, and large-urban hospitals). As noted in recommendation (3b), various coding choices for candidate attributes should be explored and evaluated.

   (b) Evaluate the policy and statistical implications of replacing the single Gaussian prior distribution for the hospital-specific random effects by a more flexible class of distributions (see appendix H).

   (c) Consider supplementing posterior mean estimates with histogram estimates. These report the distribution of the SMRs with appropriate location, spread and shape (see appendix I).

5. *Readmission rates*

   Evaluate, modify and implement the method for assessing readmission rates proposed in section 7.

6. *Model assessment*

   Evaluate augmented approaches to model assessment (see section 8).

7. *Enhance reporting*

   CMS should enhance its reporting to further emphasize uncertainty and improve interpretation. The committee suggests enhancements such as: using exceedance probabilities; juxtaposing a histogram of the patient-specific risk estimates for each hospital with a histogram of the national distribution, or of the distribution for a relevant group of comparator hospitals to clarify important between-hospital differences (see section 9).

8. *Promulgate standards of conduct and communication*

   (a) Develop and communicate standards of practice for data collection, analysis and reporting for adoption by those conducting hospital comparisons.

(b) Implement a transparent, continuous process of examining the consequences of the many, often independent, analytic choices made to ensure that what is done is as straightforward and accessible as it can be, consistent with meeting well-articulated standards for a "well-performing" quality reporting system.

9. *Transfer technology to other CMS evaluations*

   The statistical and policy issues considered in this report operate in the broad array of CMS performance measures and are relevant regardless of disease condition or process measure. Therefore, CMS should broaden its evaluations to domains other than assessment of thirty-day, post-discharge mortality and readmission rates. However, the specific choices may depend on context. For example, dialysis centers may all report a sufficient number of events so that a fixed-effects rather than a random-effects approach can be used in developing the national-level model and the SMRs.

1 Introduction

The Centers for Medicare and Medicaid Services (CMS) has a congressional mandate to evaluate hospital performance using risk-standardized mortality rates, other process-of-care outcomes, and risk-standardized readmission rates. These legislative requirements and associated challenges are not unique to the CMS. As Spiegelhalter et al. (2012) note, increasing availability of clinical outcome data and increased concern for accountability, "has led to an increased focus on statistical methods in healthcare regulation" for the three principal functions of "rating organizations, deciding whom to inspect and continuous surveillance for arising problems."

The CMS, through a subcontract with Yale New Haven Health Services Corporation, Center for Outcomes Research and Evaluation (YNHHSC/CORE), supported a committee appointed by the Committee of Presidents of Statistical Societies (COPSS) to address statistical issues identified by the CMS and stakeholders about the CMS approach to modeling hospital quality based on outcomes. In the spring of 2011, with the direct support of YNHHSC/CORE, COPSS formed a committee comprised of one member from each of its constituent societies, a chair, and a staff member from the American Statistical Association, and held a preliminary meeting in April. In June, YNHHSC/CORE executed a subcontract with COPSS under its CMS contract to support the development of a White Paper on statistical modeling. Specifically, YNHHSC/CORE contracted with COPSS to "provide guidance on statistical approaches ... when estimating performance metrics,"and "consider and discuss concerns commonly raised by stakeholders (hospitals, consumer, and insurers) about the use of hierarchical generalized linear models in profiling hospital quality." The committee convened twice, in June and August of 2011, and exchanged a wide variety of materials electronically. To ensure the committee's independence, YNHHSC/CORE did not comment on the report's findings, and CMS pre-cleared COPSS' publication of an academic manuscript based on the White Paper.

The ability to estimate *hospital performance* accurately based on patient outcome data relies upon several factors. Most basically, the outcome needs reflect something that is directly affected by the quality of hospital care. Beyond this, however, there are a number of important data and analytic considerations: (1) Data must be available and used to *adjust for* differences in patient health at admission across different hospitals (case-mix differences). These adjustments are required to ensure that variations in reported performance apply to hospitals' contributions to their patients' outcomes rather than to the intrinsic difficulty of the patients they treat. Of course, performance of the adjustments depends on the type and quality of available data, and in the CMS context data have been validated against medical chart information. (2) In distinct contrast to the previous point, reported performance should not *adjust away* differences related to the quality of the hospital. For example, if "presence of a special cardiac care unit" is systematically associated with better survival following a heart attack, a hospital's reported performance should capture the benefit provided by that unit and as a consequence such hospital-level attributes should not influence the

risk adjustment. (3) The reported performance measure should be little affected by the variability associated with rates based on the small numbers of cases seen at some hospitals.

These desired features of a performance measure require the development of a statistical model to adjust for facility differences in case-mix (a risk-adjustment model), and the use of an analytic method that reports sensible rates for all facilities, including those with few observed outcomes. In this report, the committee discusses issues in producing a credible risk-adjustment model, but focuses primarily on methods for partitioning the remaining variation into that (a) associated with variation within hospitals for a homogeneous group of patients, and (b) produced by between-hospital variation. Successful partitioning will avoid, to the extent possible, misclassifying hospitals. But even the best methods cannot completely prevent high-quality hospitals from being incorrectly characterized as having relatively poor outcomes or failing to identify the shortfalls of some lower-quality providers, especially if they are low-volume, i.e., with few patients for the procedure in question. Indeed, all statistical procedures have such operating characteristics and these must be evaluated in the context of a specific application.

This report addresses both broad and more technical statistical issues associated with mortality and readmission hospital measures. The content principally responds to the extended debriefing provided by Lein Han, Ph.D. regarding CMS's concerns (see appendix D). Subsequent sections identify key goals and relevant approaches, outline current approaches, identify possible modifications, and recommend study of the potentially most important, high leverage of these.

The committee concludes that the current CMS approach is effective, but that even within the current framework, refinements that have the potential to improve performance and credibility of the assessments should be considered. Model enhancements include use of more flexible models for case mix adjustment; broadening the class of distributions from the current Gaussian family used in the hierarchical, random effects model; evaluation of the effectiveness of current outlier detection methods; and consideration of producing an *ensemble* of hospital-specific Standardized Mortality Ratios that accurately estimates the true, underlying distribution of ratios via a histogram. In addition, within the current framework, the CMS should consider augmenting reports to improve communication and to provide additional cautions regarding interpreting results.

The decision on incorporating hospital characteristics, especially hospital volume, is crucial and relates to both the risk adjustment and stabilization components of the hospital evaluation process. All stakeholders agree that risk adjustments should not reflect hospital characteristics, but their use in reducing confounding of the case-mix/risk relation has been advocated. More contentious is their use in the component of a model that stabilizes estimates. Statistical models are available for each of these operations, and the committee presents options. The ability to develop and implement such models has never been in question, but the advisability of broadening the adjustment model from an overall shrinkage target to hospital-attribute-specific determined targets has generated

considerable debate. The committee provides information on the conceptual and technical issues associated with such models, and identifies studies and policy choices that should precede a decision on this issue.

2 Issues in Statistical Modeling of SMRs

The CMS needs to quantify the following mandate:

> "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?"

There is an explicit comparison here between the hospital at issue and a counterfactual hospital at the national level, because it is very unlikely that there is another hospital with exactly the same case-mix. Indeed, the CMS warns that one should not compare hospitals. In this regard, it is unfortunate that the CMS website reporting quality ratings is titled "hospital compare."

Krumholz et al. (2006) discuss several factors that should be considered when assessing hospital quality. These relate to differences in the chronic and clinical acuity of patients at hospital presentation, the numbers of patients treated at a hospital, the frequency of the outcome studied, the extent to which the outcome reflects a hospital quality signal, and the form of the performance metric used to assess hospital quality. Additional issues in developing, implementing, and reporting results from hospital profiling and readmission evaluation include: attaining consensus on goals, respecting the observational nature of available data and its hierarchical structure; producing valid and effective case mix adjustments; reporting credible point estimates for the hospital effects associated with patient outcomes including readmission rates; smoothing to deal with unstable estimates; and addressing the challenges in validly interpreting and effectively communicating results.

The performance measure reported by the CMS is the stabilized, indirectly risk-standardized, hospital-specific death rate (see appendix E.1). The basic estimate is,

$$\begin{aligned} \text{Standardized death rate} &= \frac{\text{Observed \# of deaths}}{\text{Expected \# of deaths}} \times (\text{the national-level death rate}) \\ &= \text{SMR} \times (\text{the national-level death rate}). \end{aligned}$$

Thus

$$\text{SMR} = \frac{\text{Observed \# of deaths}}{\text{Expected \# of deaths}}. \tag{1}$$

The denominator of equation (1) is the result from applying a model that adjusts/standardizes for an ensemble of patient-level, pre-admission risk factors, rather than only demographic factors such as age and gender as is typical in epidemiological applications. The statistical issues arising in the estimation of the standardized death rate and the SMR in the CMS assessments are identical because the latter is simply the hospital-specific value divided by the expected number of deaths computed from the national risk model. Due to the calibration provided by the SMR (SMR =

1.0 always implies typical performance relative to the national standard for the types of patients treated at the hospital), the committee focuses on it throughout this report.

2.1 Calibration to a hospital-attribute specific standard

This report focuses on profiles relative to the overall national standard because that is how the committee interprets the CMS's mandate. Indeed, all discussion in this report related to including hospital-level attributes addresses issues in how to stabilize basic estimates; not on changing the standard to which hospitals are compared. However, standardization to a comparator (a denominator) other than the overall, national standard is easily available by stratification on hospital types with separate analyses in each stratum or by generalizing the current case-mix adjustment models to include hospital attributes. For example, using a stratified approach, one might develop one SMR for community hospitals and another one for teaching hospitals. In each case an SMR = 1.0 would indicate typical performance *for a given case-mix and the specific hospital type*, and would not necessarily indicate compatibility with an overall, national standard.

2.2 Profiling versus decision-making

Two, distinct goals can be considered when evaluating mortality outcomes:

> *Profiling:* "How does this hospital's mortality for a particular procedure or condition compare to that predicted at the national level for the kinds of patients seen for that procedure or condition at this hospital?"
>
> *Decision-making:* "Given my medical status and needs, to which hospital should I go for a particular procedure or treatment of my condition?"

Though the decision-making goal is important, it is not the topic the CMS tasked the committee to address.

The profiling goal: One can address the *profiling goal* by developing a valid national-level model for the probability of death as a function of patient-specific attributes, using this national-level model to compute the expected number of deaths for a hospital's mix of patient-specific attributes (its case-mix), and comparing the actual number of deaths to this expected value, producing the Standardized Mortality Ratio (SMR). Profiling entails aggregation over a mix of patients. Thus, while a hospital with a low SMR (indicating lower than expected mortality) might be a good choice for a specific patient, it may be that the hospital performs relatively poorly for that kind of patient. Similarly, a hospital with a relatively high SMR, indicating higher than expected mortality, might perform well for that patient even though its average performance over a mix of patients is relatively poor. In addition, if the case-mix adjustment is inadequate, the playing field for hospitals will not be level, especially for procedures with a complex case-mix. Some hospitals treat only relatively easy cases while others specialize in difficult ones and any analytical approach needs to take this differential into account when computing expected deaths.

The decision-making goal: These characteristics of the SMR help differentiate the profiling goal from the decision-making goal, which is best addressed by developing a rich model for the probability of death as a function of patient-specific characteristics plus all relevant hospital characteristics. An effective model would include main effects and interactions among patient-level attributes, among hospital-level attributes and between these two sets of attributes. For example, if "presence of a special cardiac care unit" is associated with survival, then it would be appropriate to include it in such a model.

To clarify the difference between profiling and decision-making, consider two facilities treating patients having a particular medical condition. Hospital A sees primarily low-risk patients for whom the expected death rate is 6%, while hospital B is a tertiary referral hospital with most patients being extremely sick, and that collectively have an expected death rate of 16%. Suppose that both facilities perform *as expected*; that is, the observed death rates are 6% and 16%, respectively, and that the case-mix adjustment is correct. Both facilities will be profiled with an SMR = 1.0. These calculations send the right message for many purposes, but are not designed to address the decision-making goal because they were not designed to provide information about whether hospital A might be a better choice than hospital B for a particular kind of patient.

3 Components of a Profiling Approach

Developing an estimator of hospital-specific performance based on outcome measures, for example in computing the SMR (see equation 1), requires several strategic decisions. They relate to the types of covariates included in the model (patient and hospital); the statistical model for the outcome (probability of the outcome and the relation between the probability of the outcome and the covariates); and calculation of the hospital performance measure. In this context, the committee considers issues associated with case mix adjustment, low information and the observational study context. Many of these issues translate directly to assessment of readmission rates and we consider issues of particular pertinence to readmission rates in section 7.

The following components are essential building blocks for a valid approach to estimating hospital-specific SMRs.

- Respect the probability process and hierarchical structure of the data.
- Develop an effective case-mix, risk adjustment so that to the extent possible with available data, the expected number of events produced from the national-level model is free of patient-level influences, producing as level a playing field as possible.
- Stabilize the basic SMR to improve estimation and prediction performance.

The principal issues and approaches associated with each component are outlined below.

3.1 Respect the probability process and hierarchical structure of the data

Because patient-specific outcomes are binary (e.g., a death indicator), a Bernoulli model operating at the patient level is appropriate. Risk adjustment and stabilization should respect this model and thus logistic regression is a suitable approach for including the effects of patient-level attributes. Alternatives to the logistic include the probit and other links. With flexible modeling of covariate influences, each would produce a valid risk adjustment and there is no reason to replace the logistic by another link function. Furthermore, because some hospitals for some conditions have either a small number of patients or a small number of events, it is important to use the core, Bernoulli model to represent stochastic uncertainty rather than an approach that would be valid only for large hospital-specific sample sizes.

In this setting, patients are nested within hospitals so that their outcomes may be correlated due to receiving care from providers in the same hospital. While an individual may contribute more than one admission at more than one hospital for the same procedure or condition (of course, for the death outcome there is only one event), CMS does not use linkage at this level and thus patients are effectively nested within hospital. A hierarchical model is most appropriate for respecting this nesting structure, and section 4.3 contains additional information on this topic.

3.2 Develop an effective case-mix, risk adjustment

The evaluation process must be based on an effective case-mix, risk adjustment so that to the extent possible with available data, the expected number of events produced from the national-level model is free of patient-level influences, producing as level a playing field as possible. Though one might wish to have additional information of patient attributes and clinical severity, even with currently available data the CMS should evaluate whether a more flexible case-mix adjustment model will improve performance. Most important is evaluating when one should augment the model to reduce potential confounding by hospital of the patient-attribute/risk relation. The committee discusses approaches in section 4.

Patient attributes are of the three types, measured and accounted for, measurable but not accounted for, and attributes that are difficult or impossible to measure. All agree that risk adjustments should include pre-admission medical conditions, but whether or not to include demographic attributes is a policy decision, one with clear consequences. For example, if outcomes for minority group patients are generally less good than for the majority group and race is included in the risk adjustment model, then hospitals that treat a relatively large number of minority patients will get credit because the expected number of events will be larger than if race were not in the model. Therefore, including race would give a hospital credit for treating this category of patients. Of course, the reverse is true; omitting race from the risk adjustment model may be seen as penalizing such hospitals. Variation in outcomes that is due to patient attributes that are omitted from the adjustment (or adjustments that use them inappropriately) is absorbed by the hospital-specific

effects that are used to compare performance and omissions of this type may not sufficiently "level the playing field" when comparing hospital performance.

3.3 Stabilize the basic SMR

Most, but not all, stakeholders agree that the method used to estimate the SMR requires some form of stabilization, at least for public reporting (administrative reporting and quality assurance and quality control reporting may not need it). When the number of events used in a direct estimate of an SMR is small, the estimate is unstable, with a relatively large standard error or coefficient of variation (the standard error divided by the estimate), and its statistical performance is poor. In the following, "directly estimated" quantities refer to estimates based on hospital-specific, observed mortality events divided by the expected number of events obtained from a logistic regression model.

The hierarchical, Bayesian formalism provides a valid and effective approach to this stabilization goal, though other formulations are possible. The approach posits a prior distribution for hospital-specific, random effects and a "data model" (the core logistic regression) that is a function of these effects and patient-level pre-admission attributes. The Bayesian formalism adopted by CMS produces the posterior distribution of the hospital-specific random effects and uses this distribution to produce the reported SMR estimates, uncertainties and other descriptors. The approach stabilizes estimates because directly estimated SMRs (see equation 1) are shrunken toward the prior mean by an amount that depends on the standard error of the estimate. Relatively unstable estimates are adjusted to a greater extent than are relatively stable estimates.

The use of the Bayesian formalism is by no means a panacea. All models require careful specification and rigorous evaluation. Model performance depends on correct specification of the "data model" (e.g., the core logistic regression). Validity of the hierarchical approach also depends on the form of the distribution for the hospital-specific random effects and the summary used to do the stabilization. The current CMS approach uses a Gaussian distribution for hospital effects measured in the logistic scale, and the committee identifies generalizations of this approach in appendix H.

These somewhat technical issues are important, but essentially not controversial. The key issue of contention relates to whether the shrinkage target should depend on hospital-level attributes, with volume being the variable of greatest concern (see section 6).

3.4 The current CMS approach

The current CMS approach (see appendix E.1 and equation 3 for details) provides a valid and effective approach that respects the data structure. It incorporates the primary components of variance at the hospital and patient levels. It implements risk adjustment for case-mix and stabilization of the estimated SMR by shrinkage toward the national mean value using all patients from all hospitals (the national mean produces SMR = 1.0). The method's basic construct however

does not directly accommodate allowing hospital-level attributes to determine shrinkage targets for stabilizing estimated SMRs.

The committee concludes that the form of the current model, while effective, potentially requires augmentation to reduce confounding of the patient-attribute/outcome relation in developing national-level expected events. Also, to a degree, the form of the current model gets in the way of constructive debate on inclusion of hospital-level attributes in the stabilization step. Indeed, its seamless integration of case-mix adjustment and stabilization (a major virtue) tends to obscure their essentially separate roles. Clarity in this regard will by no means end the debate, but it can focus the debate on the correct issues. Therefore, in appendix F the committee presents two modeling approaches that permit shrinkage targets that depend on hospital attributes while preserving the national-level referent.

3.5 Fixed-effects and random effects models

An alternative approach uses fixed effects (FE) models to profile facilities; this approach is currently being used by CMS in their Dialysis Facility Reports. See,

`http://www.dialysisreports.org/pdf/esrd/public/SMRdocumentation.pdf`

The hospital-specific intercepts are assumed to be fixed parameters to be estimated individually rather than random parameters (effects) that are assumed to be sampled from a probability distribution (the prior). The magnitude of the estimated variance of the prior captures the unexplained, between-hospital variation. One of the major distinctions between random effects (RE) and fixed effects (FE) models is the degree of attention to the structure of the correlations between the observed and unobserved variables, correlation that can confound the patient-risk relation. In the basic RE model with an intercept-only probability distribution, there is no explicit attention to the potential correlations; in effect the unobserved variables are assumed to be uncorrelated with the observed variables. In a FE model, saturated at hospital level (there is an intercept for each hospital) there are no such assumptions, so the model provides complete adjustment for these potential correlations.

However, the committee emphasizes that there is a continuum between the basic, RE model and the saturated FE model. By augmenting the probability model used in the RE approach to include hospital attributes, the model occupies the middle ground between basic RE and full FE. Including these attributes "explains" some of the between-hospital variation that was unexplained in the basic model and the estimated variance of the prior is reduced. Including additional hospital-level attributes or interaction terms until the model degrees of freedom equal the number of hospitals results in the FE model. Because there is no between-hospital variation left to explain, the estimated variance of the prior is zero.

Therefore, there is a continuum of approaches to account for correlations and adjust for potential confounding. One can go "all the way" with the FE approach, or go "part-way" by building a RE model that includes adjustment for hospital-level attributes, but does not use up all hospital-level

degrees of freedom. If all hospitals had sufficient, stand-alone data, then, subject to the form of the model being correct, the FE approach will ensure successful adjustment for potential confounding. However, if there are small numbers of patients or events in some hospitals, the FE approach can be biased and in any case produces higher variability. The intercept-only, RE model is consistent but only under the assumption of no correlation of the random intercepts with patient case mix. The more general, hierarchical random effects (RE) model with appropriate augmentation by a hospital-level regression component produces estimates with the smaller MSE by trading off some bias for lower variance relative to the FE approach. The bias can be controlled by including sufficient hospital-level covariates and thereby producing very low correlation between the residual hospital-level variance components and patient-level case-mix.

Wolfe & Kalbfleish (unpublished) have compared the properties of FE and the basic RE model for the purpose of profiling facilities under various conditions. When there is correlation between patient risk factors and hospital characteristics, such as when sicker patients are admitted systematically to either better- or worse-performing facilities, then basic RE estimates are biased, have larger MSE, and have less ability to detect exceptional facilities. Although RE have lower MSE on average, they showed that MSE is larger for centers with exceptional performance, in other words, those whose performance can be distinguished from the national mean. FE methods have substantially higher power to detect outlying facilities than do RE models. When there is unmeasured confounding between patient risk factors and the hospital intercepts, the basic RE model does not provide accurate case mix adjustment.

The Wolfe & Kalbfleish research shows that one needs to move beyond the basic, RE model. However, the committee notes that FE approach produces larger standard errors than RE estimates, leading to wider confidence intervals. RE models allow for some residual correlation, possibly some residual confounding , but reduce variability. Striking an effective variance/bias trade-off is a central tenet of all statistical modeling, and the committee recommends that the CMS augment its current model to include hospital-level attributes with the goal of producing a variance/bias trade-off that generates case-mix adjustments with very good predictive performance.

3.6 Possible modeling approaches and the low information context

One might conceptualize the modeling task in several different ways, for example, through aggregated models at the hospital level. The primary problem with this approach is the inability to control for patient- level risk factors. Patient level models make intuitive sense because they permit optimal control for patient-level confounding and allow for inferences to the outcomes of individual patients, since this is the level at which doctors make decisions and where the outcomes of those decisions occur.

Hospital assessments of the sort that lie at the core of the CMS mandate occur in a relatively low information context. For both mortality and readmission, statistical information depends directly

on the number of events more than on the number of patients. Even hospitals with a large caseload may have few deaths or readmissions. Of course, statistical information is generally smallest for low-volume hospitals. Due to the large number of patient attributes and values for them, however, even at the national level some form of smoothing/stabilization is needed to support case-mix adjustments. Smoothing and stabilization are needed for both the numerator and denominator, when producing hospital-level, estimated SMRs or readmission rates. In effect, stabilization is accomplished by replacing the observed number of events by a stabilized value, often described as "borrowing strength" across hospitals.

The patient-level modeling strategy becomes somewhat more complex when determining which hospital-level characteristics to include in the model. How should one deal with the volume of specific procedures, hospital location (e.g., urban vs. rural, or Southeast vs. Northeast), hospital ownership and management (some hospitals are part of hospital systems which affects who goes where for specific procedures), academic teaching status, number of beds and case-mix, whether procedures are elective or emergency-based, etc.? The committee notes that the myriad possible combinations of hospital characteristics prompts the need to confront data sparseness, including the need for smoothing/stabilization in both the numerator and denominator of the SMR in expression expression (1). Section 5 discusses these issues in greater depth.

3.7 Possible data limitations

The information CMS uses to risk-adjust comes from billing claims that characterize the diagnoses observed or procedures performed during the entire hospital admission and generally fail to adequately characterize patient severity. While diagnoses present on admission (POA) are available, POA reporting accuracy varies by hospital characteristics (Goldman et al., 2011). Furthermore, for both medical and surgical admissions, although there is a code to indicate if it is emergent (versus elective) these codes may not be sufficiently reliable to support computing separate SMRs for each type of admission. As another example, many aspects of decision-making associated with the flow of patients to specific hospitals (e.g., associated with health insurance or doctors' practices) simply are not recorded. The committee notes that the CMS mortality and readmission models based on billing data have been validated against medical records data.

Hospital evaluations, whether those done by CMS or by external researchers, use observational data. Except in a relatively small number of true experiments, no one randomizes patients to hospitals for care; patients or their physicians choose a hospital. Choice depends on medical condition. For example, acute MI patients are taken to the nearest hospital, but elective surgery patients select a hospital based on various factors, such as surgical volume, physician recommendation, or word of mouth. Case mix can vary widely among hospitals. For example, coronary artery bypass graft (CABG) surgery is generally performed in large medical centers whereas community hospitals generally treat less complicated conditions and perform less complicated procedures. Even among hospitals performing CABG, some may specialize in treating the most complex and risky patients,

and others, the most routine cases. Because CMS compares hospitals relative to their own case mix, comparing SMRs between hospitals can be an "apples/oranges" exercise, and considerable caution is needed. Indeed, the CMS SMRs are produced to assess a specific hospital relative to a counterfactual population, and between-hospital comparisons are not valid unless there is a near match to the distribution of patient-specific risks.

There may well be hospital differences in coding practice and documentation. For example, for a similar set of patients, diabetes may be more frequently noted in some hospitals. More generally, hospitals may engage in up-coding and the induced differences in case mix will favor some hospitals. Hospitals differ in coding practices, for example with teaching hospitals documenting more than community hospitals (Iezzoni, 1997, 2003). Also, hospitals that engage in sophisticated up-coding can make their patients look sicker than the same patients would at a "typical" hospital, causing their case-mix-adjusted performance to look better than it otherwise would.

3.8 Endogenous and exogenous hospital attributes

The hierarchical logistic model that is the focus of this report is largely a descriptive model, but it is also imbued with a causal interpretation. Indeed, adjusting for risk has a causal interpretation. One could develop a more formal causal framework, usually associated with econometric simultaneous equations models, in which some variables are exogenous (the causes of effects) and also endogenous (the effects of causes). An interesting and perhaps important question is whether the hierarchical logistic model for outcomes can be viewed as a *reduced form* equation for such a simultaneous system and the consequences thereof regarding inclusion of hospital-level attributes. As discussed in section 6, there are implications of such causal thinking in the context of the hierarchical models employed by the CMS.

4 Case-Mix Adjustment

In developing a case-mix adjustment, CMS needs to establish a standard for what outcome is expected for a hospital's patients, strictly as a function of the patients' characteristics upon arrival (POA), but in a modeling framework this reduces potential confounding of the patient-attribute/risk relation by hospital. Nothing that is "on the causal pathway" that happens after patients arrive at the hospital should be used to establish the "denominator" for that comparison. The committee provides the following discussion to highlight the strategic and tactical issues in adjustment and stabilization. The discussion is not intended to incorporate the full details of a valid approach. See appendix E for the specifics of the current CMS approach and appendix F for enhancements.

Differences in hospital *case-mix* is one of the most widely-studied topics in the field of health services research (Iezzoni, 2003). For the purposes of assessing hospital quality, risk factors generally meet two criteria for inclusion in performance models. First, they must characterize the patient's health

at admission. This requirement ensures that information that reflects how patients are handled *post-admission* does not become confounded with or masked by hospital quality, the primary quantity of interest. For example, while cardiogenic shock may be highly predictive of mortality when measured after admission, the occurrence of shock may be a consequence of poor hospital care coordination. Inclusion of this covariate would give a hospital a "pass" for such patients. Second, patient-level and hospital-level characteristics require careful justification for inclusion. For example, there is debate surrounding the inclusion of socio-demographic characteristics in hospital quality assessments, with some arguing that their inclusion may mask disparities and inequities in quality of care (Blumberg, 1987; Iezzoni, 2003; Krumholz et al., 2006).

For each medical or surgical condition, and for each admission in the database, the CMS data consist of a 30 day, post-discharge death indicator, a list of disease conditions and other patient attributes (such as age and sex), and the admitting hospital with its attributes. Patient outcomes are attributed to a hospital and variation in these outcomes can be attributed to differences in patient case-mix, random variation in outcomes amongst patients with the same characteristics treated in the same hospital, and hospital-level variation that remains after accounting for these patient-level factors. Case mix adjustment attempts to account for the patient-level factors, using the remaining hospital-level variation to compare hospital performance related to, for example, practice patterns and hospital resource availability.

While hospital attributes must not be used to "tune" a risk adjustment to a hospital type, if hospital effects are correlated with case mix factors (which is inevitably the case) the estimated case mix coefficients will be biased and the national-level risk adjustment may not accurately consolidate the "rolled-up" patient-specific risks. Therefore, the case-mix risk adjustment model should be structured to reduce confounding by these correlation effects. As outlined in,

`http://www.dialysisreports.org/pdf/esrd/public/SMRdocumentation.pdf`

adjustment to reduce confounding is an accepted approach in other facility assessment settings and the committee recommends that serious consideration be given to such adjustments in the CMS context. However, when the number of patients or events is small, the saturated, fixed-effects approach cannot be used. In appendix F.1 the committee proposes a Bayesian hierarchical model that accomplishes the adjustment. The approach builds a model for hospital-level attributes when developing the risk model, but risk predictions are for a "typical" hospital and so these attributes do not otherwise play a role in producing the expected value for a hospital.

The foregoing discussion highlights that inclusion or exclusion of an adjusting variable and its role in a model all influence the question being asked, for example the reference to which a hospital's performance is compared. Candidate factors divide into three categories:

1. *Pre-admission, patient-level health attributes (i.e., case-mix):* All stakeholders agree that using information from this category is not only appropriate, but necessary. To produce a fair assessment, predictions must account for patient-level, upon-admission characteristics that associate with outcome. The committee notes that even here several issues must be

considered, for example whether to include a patient's race in the model. These and other such decisions are primarily ones of CMS policy and not of statistical practice.

2. *Post-admission patient attributes including events that coincide with or might be the result of care (e.g., in-hospital infections or patient length-of-stay):* All stakeholders agree that including this category of information is inappropriate. Adjustment should not be made for post-admission events, because they are on the pathway to the outcome and adjusting for them would reduce the magnitude of the hospital effects.
3. *Pre- or at-admission, hospital attributes (i.e., presence of cardiac catheterization laboratories):* There is disagreement on whether to include information from this category. For example, should hospital location, number of beds, condition volume, etc. be included in the model used to stabilize estimated hospital effects?

4.1 Inclusion of hospital-level attributes in the risk model

It is clearly the case that hospital-level attributes from category (3) such as volume for different procedures, location (urban, rural), ownership and management (for-profit vs not for profit), mission (teaching, community) status should not be used when computing expected deaths when "rolling up" the patient-specific probabilities in the national model that is used to produce the denominator for an SMR. Doing so would set a hospital attribute specific standard and an SMR = 1.0 would indicate that the hospital has performance that is typical for hospitals with the same case mix *and the same hospital-level attributes*. For example, urban hospitals would be compared to urban, rural to rural, and the SMRs would not quantify performance of rural hospitals relative to urban. While such comparisons might be of interest for some purposes, they would defeat the primary purpose of national-level profiling of enabling stakeholders to compare all hospitals to the national standard.

There remain two other potential roles for category (3) attributes in national profiling:

1. Reduce confounding by hospital in the patient-level risk model. Appendix F.1 reports on a model to accomplish this goal.
2. Enrich the model for estimating hospital effects beyond use of the current, intercept-only, hierarchical logistic regression model that stabilizes hospital effects (and thereby estimated SMRs) by shrinkage to a single target. Section 5.2 provides additional discussion of this issue.

4.2 The national-level model

The SMR measures the ratio of what has been observed to what is expected in a particular hospital for the same patient case mix using a national-level model. To estimate this ratio, a model for the national level probability of death as a function of patient-level attributes is needed. To obtain the expected number of deaths for a specific hospital, a probability of death (or readmission) is computed for each of its admissions and these are summed. In developing the national model, because the patient-specific outcomes of death or readmission are binary, logistic regression treating the probability of death as a linear logistic function of variables has become the preferred approach in developing a case mix model from national data, both for CMS and for most other health services

researchers. Successful case-mix adjustment depends on building an effective regression model for the patient-level attributes in the context of hospital-level variation. The full armamentarium of statistical modeling is available to accomplish this task.

Hierarchical versions of logistic regression are a natural component of such models, because patient admissions are clustered within hospitals. In building the national-level model, the hierarchical structure of the data should be accommodated to the degree possible, and the model for the probability of death must be adjusted for relevant patient-level information, but not for hospital-level attributes. The standard logistic regression should include a sufficiently rich model for the influence of at-admission, patient attributes and to accommodate the clustering. To deal with possible confounding, hospital-specific intercepts are necessary. Standard choices for these intercepts are either *fixed* (there is a separate, explicit intercept for each hospital, often called "dummy" or indicator variables) or *random* (the hospital-specific intercepts are modeled as a random sample from a probability distribution).

As discussed in section 3.5, the fixed-effects approach does provide input to a national-level, risk adjustment model based on the association of patient-attributes with the probability of death. However, by saturating the hospital-level model, it cannot incorporate the stochastic effects of clustering. Therefore, the committee recommends use of the hierarchical, mixed-effects approach for the national-level model of mortality outcomes (i.e., replacing the hospital-specific, fixed effects by an assumption that these effects are random variables drawn from a distribution of such effects). Importantly, the model must also adjust for hospital-level attributes to reduce potential confounding of the patient attribute-risk relation. Such models allow for both between- and within-hospital variation, but do not require *à priori* that they exist. The estimated variance of the distribution considered to generate hospital-specific effects measures the extent to which they exist above and beyond that associated with the hospital-level attributes included as covariates. Such a components of variance perspective is shared by a number of different families of statistical models and leads naturally to a multi-level hierarchical model perspective rather than a standard logistic regression.

4.3 Stabilization via hierarchical models

The broad CMS goal is to judge a hospital's performance based on what happens to the patients it actually admits with the aim of determining the hospital's influence on outcome. Therefore, profiling models should adjust for important patient risk factors and incorporate the clustering of patients within the responsibility of the hospital. Hierarchical regression models, including hierarchical generalized linear models, respect this data structure and have other benefits. They explicitly quantify both intra-hospital (patient-level) and inter-hospital-level variation. The estimated hospital effects (like hospital "signatures") quantify the case mix adjusted residual effects of hospital quality on patient outcomes, measured as deviations from the national mean. Variance component modeling induces within-provider correlation (an intra-class correlation, ICC) in that the outcomes of patients within the same hospital tend to be more similar than if the same patients were treated

in different hospitals. It is possible that there are no hospital quality differences, that is the chance that a patient experiences an event after being treated is the same regardless of the hospital. In this case the inter- hospital variation would be zero as would the ICC, and all hospital effects would be zero, implying no difference from the national mean.

4.4 Stabilization via hierarchical modeling

Stabilizing estimates is a fundamental statistical practice, in all cases employing some form of borrowing strength or information. Regression models and local averaging (e.g., LOESS Cleveland and Devlin, 1988) are statistical mainstays of the statistical toolkit for this purpose, as are more complex methods involving splines (Wang, 2011), wavelets (Morris and Carroll, 2006), and functional data analysis (Crainiceanu et al., 2011). Hierarchical models with shrinkage towards a regression surface (Carlin and Louis, 2009; Gelman et al., 2004) provide a model-based way to produce estimates that are close to the inferential target (producing a relatively small mean squared error) by stabilizing variance while retaining sufficient alignment with the target. Standard regression approaches stabilize by using the model-based predictions and the estimates; that is the direct estimates (for example, the data points) are moved all the way to the regression surface. Hierarchical models stabilize by moving direct estimates (in our context the $\hat{\beta}_{0i}$ in equation 2 of appendix E.1) part-way towards the regression surface, thereby retaining some of the hospital-specific signal.

Hierarchical models produce standard errors that incorporate the components of variation within and between hospitals. They also provide a framework for stabilizing estimated hospital-specific effects by shrinkage towards a central value. Stabilization dampens the regression to the mean effect, the phenomenon wherein hospitals found to be at the extremes in one year subsequently become less extreme, thereby stabilizing a sequence of assessments. This form of stabilization also reduces the influence of chance in the observed variation among providers. Theory and practice have shown that hierarchical models, carefully applied, produce estimates and predictions with excellent operating properties, obtained by trading-off prediction variance and bias to produce lower expected squared deviation (referred to as mean squared error) from the true, underlying relation. They have proved to be successful in a wide variety of application settings, e.g., see Bishop et al. (1975); Carlin and Louis (2009); Gelman et al. (2004); Normand and Shahian (2007). Fienberg (2011) and the associated discussion provide an excellent review of the role of Bayesian hierarchical models in the policy arena. The committee provides the following small sample of examples.

Teacher evaluations: Teacher-level performance is measured on the basis of student performance on standardized achievement tests (Camilli et al., 2001; Lockwood et al., 2002; Whoriskey, 2011). As is the case with patients, students are clustered within classrooms and vary in their scholastic abilities. As is the case with hospitals, classroom and school sizes range from small to large.

Employment discrimination (stopping short): In legal proceedings focused on employment discrimination, expert witnesses using regression models are typically required to leave out employer-controlled variables such as rank or title that are potentially tainted and thus would mask the direct measure of discrimination as captured by a regression coefficient for sex or race, e.g., see Dempster (1988); Greiner (2008).

Census adjustment (shrinkage and controversy): In the debates over census adjustment in the 1980s, Ericksen and Kadane (1985) proposed a regression based adjustment model that smoothed sample based adjustments using socio-economic and other variables to get small area adjustments where there was little or no direct data on adjustment. Freedman and Navidi (1986) countered that it was inappropriate to smooth data across state boundaries to produce adjustments that could be used to reallocate congressional seats among the states.

Small area estimates of income and poverty (shrinkage and little controversy): Effective estimation in small area requires stabilization of area-specific estimates while maintaining sufficient geographic focus. For example, Citro and Kalton (2000) report on the use of hierarchical models to produce estimates of income and poverty in small geographic areas. These estimates are used to allocate billions of dollars in school aid funds and to implement many other federal programs.

Automobile insurance rate making (shrinkage to balance the books): In the automobile insurance industry, there is a long, successful history of using hierarchical models to improve the predictive performance of estimated accident rates. Data are cross-tabulated into a large number of categories formed for example by age, gender, marital status, and rating region. Direct estimates are quite noisy and shrinkage improves performance. See Tomberlin (1988) for an informative example.

Healthcare regulation: rating, screening and surveillance (developments continue): Spiegelhalter et al. (2012) provide a variety of examples of goals and approaches used or being considered in the United Kingdom.

5 Inferences in a Low-Information Context

This section provides additional details on case-mix adjustment followed by consideration of issues associated with stabilizing directly estimated SMRs.

5.1 Dealing with low information in case-mix adjustment

All stakeholders agree that some form of stabilization is needed to produce case mix adjustments based on a large number of pre- or at-admission patient characteristics. The cross-tabulation approach, using for example the approximately 28 characteristics along with a binary indicator of death within 30 days of discharge, virtually always produces a small number of patients per cell. Indeed, if all 28 characteristics were binary, cross-classification would result in $2^{29} \approx 10^{13}$ cells. Dealing with this decreasing direct information as the number of cross-classification cells increases is

a long-standing statistical problem that can be addressed by overlapping but identifiable approaches including statistical modeling, direct smoothing, and aggregating. For example, in the 1960s in the context of the National Halothane Study, Bunker et al. (1969) (see also Mosteller, 2010) applied a variety of approaches to very sparse count data to predict mortality associated with the use of different anesthetics in a large number of hospitals for different types of procedures and with widely varying levels of risk. The halothane committee analyzing these data attempted to use direct and indirect standardization, various forms of data smoothing including log-linear and logistic models as well as early versions of hierarchical models. The hierarchical logistic regression models used by the CMS have directly evolved from this work and the class of approaches continues to expand (see section 8).

5.2 Dealing with low information in stabilizing estimated SMRs

The most politically and technically challenging aspect of statistical modeling in the CMS context is how to smooth or stabilize estimates when some, or even many, hospitals treat far too few cases of a particular kind for their data to provide stand-alone information as to their quality. This issue is intertwined with what to report and how to report it. For example, in a hospital where only 2 people were admitted with a heart attack, their observed 30-day mortality can only be 0%, 50% or 100%. However, the national mortality rate for heart attack admissions is about 16% and any individual hospital's true rate is extremely unlikely to lie outside the range from 5 to 25%, so it would be a mistake to report any facility's rate as 0% or 50% based on these data. But, CMS must report a value for that hospital nonetheless, and the uncertainty associated with its directly estimated observed rate will be much greater than for a hospital with a much larger patient volume.

The SMR, computed as the stabilized hospital-specific number of events divided by the national-level model predicted number of events, is the object of inference. Disagreement and contention focus on the approach to stabilizing the numerator of this ratio, not the denominator. All stakeholders agree that neither volume nor any other hospital-level attribute should influence the national-level model predicted events. However a case can be made for the number of beds and other hospital-level attributes to play a role in stabilizing the numerator. The current CMS approach stabilizes estimated SMRs via a hierarchical, random effects logistic regression model that shrinks directly estimated hospital effects towards an overall mean. Other methods are available, but methods that do not stabilize/smooth in some fashion will be less stable and very likely less accurate.

If a large amount of information were available for all hospitals and all basic SMR estimates were very stable, there would be no need for additional stabilization and these estimates could be reported along with what would be very narrow confidence intervals. However, it is common that many hospitals have few deaths and some hospitals have a small number of patients for a specific condition and so estimated SMRs are highly unstable. Though reporting SMR estimates along with appropriate confidence intervals will communicate this uncertainty, stabilization can be used to reduce variability while retaining sufficient year-specific, hospital focus, and generally improving

performance. Combining input data over several years is one approach (the CMS uses the three most recent years), but considerable estimation uncertainty remains and pooling over additional years reduces the sensitivity in identifying changes over time. Therefore, a model-based approach is needed.

Stabilizing through a form of shrinkage that considers all hospitals simultaneously, low and high volume, has the effect of producing estimates for low volume hospitals that are close to the national average (i.e., to SMR = 1.0). This phenomenon is one of the principal critiques of the current CMS approach by some stakeholders, those who argue in favor of different shrinkage targets for low and high volume hospitals, and those who argue for no shrinkage at all. Others note that use of different targets appears to run counter to the CMS mandate or that in any case extreme care is needed in evaluating the consequences of such an approach. In section 6 the committee addresses these issues in detail.

Irrespective of decisions on model augmentation, if in addition to producing hospital-specific estimates the CMS reported the actual number of cases for each hospital rather than simply noting whether this number was < 25, users would be alerted to low information and high uncertainty. The committee acknowledges that in some circumstances revealing the count might lead to privacy problems and individual disclosures (e.g., see the discussion in Fienberg, 2011). Therefore, rather than recommending that the CMS report the actual number of cases, the committee notes that the complete reporting of statistical properties, e.g., the point estimate and confidence interval, would communicate caution when the number of cases is low (see section 9).

6 Volume and Other Hospital-level Attributes

The SMR is the object of CMS inference. Some argue that if shrinking to an overall mean makes sense then so may shrinking towards an augmented regression model that could include volume. With over 4000 hospitals being evaluated, sufficient degrees of freedom are available to augment the model and still have a stable system. Appendix F describes models to accomplish the task if there is agreement on the conceptual base for doing so.

Hospital volume plays an important role in hospital quality assessments because the amount of information to assess hospital quality depends on the number of patients treated and, with event data, more particularly the number of observed events. Thus, unless the analysis includes some form of stabilization, hospital performance estimates associated with low-volume hospitals will be noisy. For example, at the national level, CMS pneumonia measures have mortality rates ranging between 9% and 20% and readmission rates between 16% and 26%. However, for a small hospital the randomness associated with sample size alone can be rather large. Similarly, the median annual Medicare Acute Myocardial Infarction (AMI) volume in 2006 was 15 patients per hospital for the 4171 US non-federal hospitals (Krumholz et al., 2011) studied for all-cause readmission following discharge for AMI. Volume varies dramatically across hospitals, however; 25% of the 4171 hospitals

had volumes greater than 59 and 25% had volumes less than 5. With a median volume of 15 AMI cases discharged alive and a national all-cause readmission rate of 18.9% for them, the probable error in estimating the unadjusted rate is at least ±19% (computed as $2 \times \sqrt{pq/n}$) for half the nation's hospitals with volume less than 15 AMI patients, and is ±11% for hospitals with a volume of 45 AMI patients. Separation of the sampling variability (that attributed to a finite "n") and case-mix effects from hospital effects on outcomes is both challenging and critical for isolating hospital quality.

A key issue is that hospital volume and possibly other attributes are both predictors of and consequences of hospital quality. Several research articles describe a volume-outcome relation, documenting the "practice makes perfect" theory when examining predictors of patient mortality for surgical procedures (Birkmeyer et al., 2002; Dudley et al., 2000; Shahian and Normand, 2003), and more recently for medical admissions Ross et al. (2010). Others have shown that hospital volume combined with surgical mortality is a strong predictor of hospital surgical mortality (Dimick et al., 2009). Some professional associations, such as The Leapfrog Group, have advocated the use of volume as a performance standard.

Low volume hospitals present a dilemma, indeed a tri-lemma! Either highly variable estimated SMRs will be set aside and not reported, or they will be reported as observed (ideally with emphasis on their uncertainty), or the estimates will be stabilized by substantial shrinkage towards either the national-level typical value (an SMR = 1.0) or a shrinkage target that depends on hospital-specific attributes. To frame the discussion, the committee first considered the *very large information context*. Imagine that for a given medical or surgical procedure all hospitals had a very high volume, sufficiently high that the expected number of deaths was also very large. In this case the SMRs produced by the foregoing method, the CMS method or other such approaches would be very close to the direct (Observed/Expected) estimates, "what you see is what you get." All hospital-level influences on outcome would take care of themselves in that the unadjusted SMR would accurately capture them. Indeed, the hospital effects would consolidate the effects of all hospital-level factors and in this "very large information" context would be equivalent to estimating a separate intercept for each hospital.

In reality, the level of information is not very large and for some hospitals is quite small. Thus, the issue of using hospital-level attributes to determine the shrinkage target is driven entirely by the low information context. These decisions primarily affect the low volume hospitals, because the adjustment of the estimates for high volume hospitals is relatively minor. To see this, consider a hospital that for a given condition has treated almost no cases. The current CMS approach will estimate that hospital's SMR as very close to 1.0, irrespective of the value of the events/patients ratio. That is, the hospital will be reported as operating approximately at the national standard. Although extreme, this situation is actually quite common. The committee considered whether the shrinkage target should be tuned to some hospital-level attributes with the target determined by national-level data.

Arguments in favor of maintaining the current CMS approach with shrinkage of all hospital-specific observed rates towards an overall national mean (and thereby shrinking observed SMRs towards 1.0) include the concern that it could be unwise to include only one hospital-level attribute, specifically volume, especially when it has a partially endogenous character.

Silber et al. (2010) provide the principal argument in favor of including volume in determining the shrinkage target. They showed that lower quality is associated with lower volume, and that the shrinkage target for stabilizing the numerator of the estimated SMR is substantially modified using a volume-dependent target. Furthermore, for the procedures they considered, including hospital characteristics in addition to volume did not substantially change the shrinkage target compared to using volume alone (because characteristics such as cardiac hospital, were related to hospital volume). This conjunction of the very attribute that produces large shrinkage also being associated with performance energizes the debate.

The case for including volume is by no means "sealed" by these results and the committee notes that three principal points require careful consideration. First, if some low-volume hospitals perform well, it would be unfair to move their SMRs too far towards a single, low-volume hospital target. To the extent that other hospital characteristics are available to identify well-performing small facilities, it would be better to include both volume and other attributes, possibly with interactions, so that the shrinkage target is better tuned to characteristics. There are thousands of hospital-level degrees of freedom available and it is unlikely that more than a small fraction would be needed to capture most hospital-attribute-related variations in outcome. Of course, models must be procedure or condition specific, adding another level of complexity.

The second principal point is that volume has a combined role as both an exogenous attribute that may be independently associated with quality but not "caused" by quality (e.g., practice makes perfect), and an endogenous attribute insofar as today's low volume could be a consequence of previously observed poor quality, and therefore, in the causal pathway between the exposure (the hospital) and the outcome (e.g., mortality). To clarify the complexities, the committee reiterates that volume not be included in the model for the expected rate, or the denominator, for exactly these reasons. However, the issues regarding inclusion in the stabilized observed rate, or the numerator, is different since the numerator is meant to represent the "best" stabilized estimate of the observed rate for comparison to the expected rate. Here, there are many reasonable choices for the shrinkage target, depending on the objectives of the evaluation. Indirectly standardized rates use the observed rate in the numerator; this rate is unbiased but may be too unstable for low-volume hospitals. By shrinking to a hospital-attribute-specific mean, hierarchical models would shrink to a value that best represents a small hospital's stabilized estimate. Though there is reported research documenting the relation between volume and quality, the committee neither endorses nor denies use of volume as a component of the shrinkage target. Rather, the committee calls for careful study of this issue to understand the consequences of volume's combined endogenous/exogenous role. There is a distinction between volume of patients for the condition studied and size of the

hospital as measured by the number of beds. Hospital size is fixed and exogenous, and thus the only issue is how one should utilize size in the modeling process.

The third argument regarding having the shrinkage target depend on volume is that "low volume" may well be a marker for an inadequate risk adjustment that disadvantages small hospitals. This failure to level the playing field is quite possible in that though low volume hospitals contribute to development and estimation of the national level risk model, they are given considerably lower weight than are the high volume hospitals. Consequently, the "fit" is generally better for the larger volume hospitals and the national level expected events better reflect the severity mix for treated patients. It is the case that the random effects approach gives low volume hospitals relatively more weight than does fixed effects approach (another advantage of random effects in the CMS context), but differentially inadequate risk adjustment is still possible. There may be unmeasured or inadequately modeled, or inadequately risk adjusted patient severity. See sections 3.5 and 4 for discussion of the relation between fixed and random effects models.

6.1 Recommendation regarding hospital-level attributes

The committee advises that additional evaluation is needed before deciding which hospital attributes should be used in setting shrinkage targets. A strong case can be made for using hospital-level attributes such as number of beds, both to reduce confounding when estimating the national-level, risk adjustment model and to determine shrinkage targets for SMR stabilization. Covariate main effects and possibly interactions need to be considered (e.g., small vs. large rural hospitals may have a relation different from small vs. large urban hospitals). However, the issues associated with use of volume in either of these components of the assessment process are sufficiently complex and contentious, that the committee recommends only that substantial evaluation is necessary. Similarly, CMS needs to assess the issue of possibly differentially successful risk adjustment. One way to conduct these evaluations is via simulation models rooted in the CMS context and with CMS data. The committee had neither the time nor the resources to pursue these investigations.

Importantly, irrespective of the decision on including hospital-level covariates, the committee recommends that the data from low volume hospitals be included in national assessments. These data are important for determining model parameters (population mean, between-hospital residual variation, hospital-level regression parameters, etc.), and the total information so provided is considerable. In addition, reporting stabilized estimates even for very small hospitals avoids the need to determine a minimum number of cases needed before reporting an SMR. Reports need to be accompanied by clear explanation of the estimation process and caveats on over-interpretation. Finally, inclusion in the national assessment allows stakeholders to see their performance, and the very fact of inclusion can improve performance.

7 Readmission Rates

Recent congressional legislation requires CMS to evaluate all-cause hospital readmission rates using an action threshold of 1.0. That is, at a specific hospital, when the ratio of the number of 30-day readmissions to the predicted number of such readmissions computed from national data exceeds 1.0, the hospital is penalized. Modeling used by the CMS to create the expected readmission rate is similar to that for mortality outcomes although there clearly are empirical differences. See Krumholz et al. (2011) for details.

The committee notes that Congress has imposed the threshold of 1.0 and that the current rules regarding an action threshold do not account for stochastic uncertainty in the estimated ratio. This lack of accounting produces unfair results for both high and low volume hospitals. For example, a ratio of 1.1 produced by 11/10 produces the same penalty as for the same 1.1 ratio produced by 110/100; a ratio of 0.9 produces no penalty (and no reward) also irrespective of the size of the denominator. Thus, a low-volume hospital with a true, underlying readmission rate of 1.1 is very likely to have an estimated rate below 1.0 incurring no penalty; a high-volume hospital with the same 1.1 underlying rate will be penalized most of the time. Similarly, a low-volume hospital with a true rate of 0.9 will frequently produce an estimate rate that exceeds 1.0; this will happen very seldom for a high-volume hospital. In summary, for low-volume hospitals in this situation there will either be high false positive rate (true rate is below 1.0) or a high false negative rate (true rate exceeds 1.0). The high volume hospitals are disadvantaged relative to the low because their false positive and false negative rates are relatively low, producing what can be relative unfairness. Of course if the true, underlying rate is very close to 1.0, then even high-volume hospitals will have a high false positive or false negative rate.

This high volatility is a problem similar to that in estimating SMRs, but has added impact because ratios that exceed 1.0 generate a penalty, but ratios that are below 1.0 do not generate a reward. To address these issues, the committee recommends that the CMS adopt an approach to reduce the volatility and improve the operating characteristics of their estimator, by reducing both the false positive and false negative rates. Implementation would require a change in legislation, but making the change is very worthwhile. As developed and applied by Diggle et al. (2007); Landrum et al. (2000); Landrum and Normand (2003); Lin et al. (2006) and others, the substantive change is to replace the "ratio > 1.0" rule by one that bases the decision on a computed probability that the true ratio is greater than 1.0. This computation depends on the Bayesian formalism with the target of inference being the true, underlying observed/expected ratio.

The approach can be based either on hierarchical, Bayes or empirical Bayes modeling similar to that used in computing SMRs or on *frequentist Bayes* (the posterior distribution is based on an uninformative prior and so the direct estimate is the posterior mean and the sampling variance is the posterior variance) In either case, with R_i the true, underlying ratio for hospital i, compute the posterior probability that it exceeds 1.0; $PE_i = pr(R_i > 1.0 \mid \text{conditional on the data})$. This

probability is an informative summary of the likelihood that a hospital's true ratio exceeds 1.0 and can be used to implement a penalty system. Two candidates are,

Exceedance probability threshold: CMS (or Congress) would select a $0 < \gamma < 1$ and identify a hospital as out of compliance if $PE_i > \gamma$. The value of γ determines the operating characteristic. Using $\gamma = 0.5$ is equivalent to determining whether the posterior median of R is above 1.0. This value might be appropriate if a hierarchical, empirical Bayes approach is used, but would be too low if frequentist-Bayes is used.

Pro-rata penalties: This approach would allocate penalties based on the likelihood that the true, underlying ratio exceeds 1.0 by applying penalties to all hospitals with the amount of the penalty depending on the exceedence probability, PE. Penalties could be the PE fraction of the full penalty; (full penalty)$\times$ PE or the PE could be partitioned, for example a PE ≤ 0.20 produces no penalty and so on.

The committee encourages that the CMS seriously consider these options, with specific focus on the fully Bayesian approach with stabilization of the observed rate (count/discharges) via shrinkage either to an overall national mean or to a shrinkage target that depends on hospital-level attributes. As for the SMR, shrinkage targets that depend on hospital-level attributes will primarily affect the low volume hospitals, because observed rates for high volume hospitals will be relatively stable and adjustments will be relatively small.

8 Model Development and Assessment

Model development and assessment must be conducted in the context of modeling goals and constraints. For hospital profiling the case mix adjustment goal is not to find the unrestricted best model for outcomes, but rather to find the best model that does not adjust for post-admission patient characteristics. As discussed in section 2, section 4.3 and elsewhere, a valid and effective model properly accounts for relevant patient-level information that is available at admission and associated with the outcome. Use of an effective case mix adjustment model produces expected values that support a fair comparison, an inadequate model will treat some hospitals unfairly. The *hospital effects* capture all that isn't associated with patient attributes but are associated with the outcome. These, along with stochastic variation produce the deviation of an estimated SMR from 1.0.

8.1 The patient-level model

Valid case mix adjustment depends on building a model that accurately predicts the probability of death (or readmission) using patient attributes. Thus, prediction of a binary outcome is the modeling goal for risk adjustment, and the full armamentarium of statistical models is available. The generally large number of patients supports use of rich and flexible risk adjustment models

produced by a combination of addition of interaction terms amongst patient-level predictors, use of generalized additive models or splines (Crainiceanu et al., 2007; Wood, 2006), classification trees, random forests and boosting (Berk, 2008; Breiman, 2001; Hastie et al., 2009; McCaffrey et al., 2004) and similar approaches. Model comparisons and evaluations include likelihood-based approaches such as AIC and BIC, and for hierarchical models DIC (Bayarri and Castellanos, 2007; Carlin and Louis, 2009; Gelman et al., 2005; Ni et al., 2010; Spiegelhalter et al., 2002). Data-based assessments include residual plots, stratification on predicted risk and computing standardized (Obs – Expected)/SD(Obs) (see Citro and Kalton, 2000, for examples), adjusted R^2, cross-validation via PRESS or related approaches using a more appropriate loss function for binary outcomes (Efron, 1978), the R^2 and C statistics (Ash and Shwartz, 1999; Silber et al., 2010), the Hosmer-Lemeshow and AUC statistics (Spencer et al., 2008), and a variety of sensitivity analyses (Kipnis et al., 2010). Prediction of a binary dependent variable (death/survival) is a classification goal. Care is needed in comparing models and judging absolute performance because classification performance as measured, for example by AUC and predictiveness (the principal goal in risk adjustment) can be quite different (Pepe, 2003; Pepe et al., 2008).

Understanding the restrictions on model development is especially important when using data-analytic evaluations (e.g., predicted versus observed events for partitions of the data by patient-level attributes) because the restrictions are likely to induce some large deviations. These can be investigated to see if enhanced patient-level modeling reduces them; if not they are likely associated with hospital-level attributes that are not and should not be included in the case-mix adjustment model. More generally, care is needed to use models that are as comprehensive and complex as is necessary, but no more so.

8.2 The hospital-level model

Assessment of the hospital level model can be undertaken using posterior predictive checks of key features of the between-hospital model. These may involve a comparison of the between-hospital observed standard deviation with that produced by posterior draws from the model.

8.2.1 Outlying hospitals

High-volume hospitals, especially those that are outlying from the cohort of hospitals evaluated can have large influence on a risk adjustment (Shahian and Normand, 2008), both with regard to the fixed effects (coefficients of patient attributes) and the random effect distribution. Standard regression diagnostics (e.g., DFFITS) and other "leave one out" approaches should accompany any risk adjustment. Hospitals identified to have large influence may need to have their influence down-weighted in developing a final model so as to preserve their status as outliers. Alternatively, hospital-level, random effect distributions with longer tails than a single Gaussian or with multiple modes also can be used to reduce the influence of outliers (see appendix H). For current approaches to outlier identification in a hierarchical modeling context, see Jones and Spiegelhalter (2011).

8.3 Model estimation

Due to complexity of the hierarchical model, estimation of parameters requires more effort on the analyst's part than when estimating a regular logistic regression model. The current CMS approach uses Markov chain Monte Carlo (MCMC) estimation to fit the hierarchical random effect parameters in the model, coupled with bootstrap replication to produce robust uncertainty assessments of the estimates. MCMC approximates the distributions of all model parameters (fixed and random effects) by sequentially sampling from conditional distributions. The approach requires use of good starting values to ensure model convergence (a similar requirement applies to other recursive methods), execution of more than one chain to assess model convergence, and designation of a lag with which to sample from the draws when computing distributional features of parameters (e.g., use every 5^{th} draw for computing parameter estimates).

Some authors have encountered convergence difficulties (e.g., see Alexandrescu et al., 2011) emphasizing the need for care and up-front work. Moreover, the capability of software packages to estimate model parameters varies; constraints such as the number of random effects, the number of covariates, and the number of observations per hospital are examples of such features.

Other approaches to modeling and model fitting such as empirical Bayes, penalized quasi-likelihood (PQL: Lin, 2007) or generalized estimating equations (GEEs) are available, some more easily communicated than the current approach. With non-linear models, implementing the empirical Bayes approach requires numerical integration which is itself best done by Monte-Carlo. PQL and GEE may perform well for some types of conditions or procedures (e.g., those with high event rates), they may break down in low volume or low event-rate situations. The committee concludes that investigation of alternative fitting methods has a low priority, but that CMS needs to improve its communication on the basic ideas behind their approach. This report provides guidance in this regard.

9 Reporting

Informative reporting coupled with appropriate cautions in interpretation are necessary for communicating complicated goals, concepts, and procedures. Furthermore, most report content must communicate effectively to a broad range of stakeholders, including the CMS, others engaged in policy development and implementation, elected officials, hospital administrators, insurers, and the general public. These groups will have different interests and levels of understanding. Some will be allowed access to information not available to others. Of course HIPAA and other disclosure protection requirements must be met.

9.1 Transparency

There is a broader issue of recommending a transparent, ongoing process of examining the consequences of the many, often independent, analytic choices made; to ensure that what is done is as

straight-forward and accessible as it can be, consistent with meeting well-articulated standards for a well-performing quality reporting system. CDC (2010) provides an excellent example of debriefing on methods and provides a website for "do it yourself" computation of SIRs related to central line acquired infection rates (CLABSIs).

9.2 Communicating uncertainty

Current reporting clearly reports results using tables and graphs that communicate point estimates and uncertainty. Details are available on the `hospitalcompare.hhs.gov` website. The website shows how many hospitals (both nationally and within a state) were found to be better, worse or no different from the national rate (and how many had too few cases to make a clear statement). See also,

```
http://www.hospitalcompare.hhs.gov/tables/hospital-oocQualityTable.aspx?hid=
220110%2c220031 %2c22010&lat=42.3380341&lng=-71.09286029999998&stype=
MEDICAL&mcid=GRP_4& stateSearched=MA&stateSearched=MA&measureCD=
&MTorAM=MORT
```

and the related graphical presentations at for example,

```
http://www.hospitalcompare.hhs.gov/Graphs/Hospital-OOCGraph.aspx?hid=
220110,220031,22010F&stype=MEDICAL&mCode=GRP_4&MTorAM=MORT
```

The committee recommends some enhancements, the most important being development and implementation of improved methods of communicating uncertainty and associated cautions. Confidence or posterior intervals and stacked confidence bands (Spencer et al., 2008; Spiegelhalter et al., 2012), are necessary and the experienced consumer will at least informally integrate the point estimate and uncertainty, but the take away message will always be the point estimates of the SMRs. If point estimates were better tempered by uncertainty, some (but by no means all) of the contention that surrounds shrinkage of hospital effects toward the national mean and thereby the SMRs toward 1.0 would be reduced.

A new reporting format proposed by Louis and Zeger (2008) might help. The idea is to emphasize that an estimate *is composed of* the point value and its associated uncertainty by *connecting them at the hip*. For example, rather than reporting that the estimate is 0.20 with 95% confidence interval (0.15, 0.26), the report would be that the estimate is ${}_{0.15}0.20_{0.26}$.

Other options include restricting reporting to confidence intervals with no point estimate, or to comparing the confidence or posterior interval for the risk-standardized death rate (or SMR) to the U.S. national risk-standardized rate (or to an SMR value of 1.0). If the interval estimate includes (overlaps with) the national value, then the hospital's performance is, "no different from the national standard." If the entire interval estimate is below the national value, then the hospital is performing better than the national standard; if the entire interval estimate is above the national

value, then the hospital is performing worse than the national standard. Hospitals with extremely few cases or events in a three-year period would generate either a very broad confidence interval or a posterior interval with the national standard far in the interior. In addition to reporting, for example, that the number of admission is < 25, this uncertainty helps to quantify the statement, "the number of cases is too small to reliably tell how the hospital is performing."

9.2.1 Threshold exceedance probabilities

In addition to numerical and graphical communication of uncertainty via confidence or posterior intervals, the committee recommends augmenting reports by (and possibly basing policy on) a summary other than the traditional point estimate with an uncertainty interval; specifically on some other features of the full SMR uncertainty distribution. Building on ideas contained in section 7 regarding re-admission rates, the committee proposes reporting *exceedance probabilities*, i.e., pr(SMR $> s$ | conditional on the data) for several values of "s" or alternatively reporting the (5th, 25th, 50th, 75th, 95th) percentiles of the SMR. Elevating pr(SMR > 1.5 | conditional on the data) or some other threshold to be the primary measure of hospital performance would force attention on something other than the center of the distribution and would effectively incorporate uncertainty. This approach would to some degree calm the intensity of the debate about shrinkage, because while the exceedance probabilities computed from the posterior distribution or from the *frequentist Bayes* approach (see section 7) are different, the differences are smaller than for the point estimates. Furthermore, if a low degree of freedom t-distribution were substituted for the Gaussian prior distribution of the between-hospital variance component in the hierarchical model, differences between the Bayesian and frequentist Bayes exceedance probabilities would be further reduced, while retaining much of the stabilizing effect of the Bayesian approach.

9.3 Encouraging appropriate interpretations

To emphasize that in general hospital SMRs should not be compared with one another (a caution that applies to all indirectly standardized rates), histograms of the patient-specific risk estimates for a hospital along with the national distribution or a set of other *relevant comparator* hospitals should be available to at least some stakeholders (see Shahian and Normand, 2008).

A bivariate display may also encourage appropriate interpretations by discouraging inappropriate comparisons. As motivation, consider that the indirectly adjusted rate is meaningless without knowing the national rate for the counterfactual hospital. A number like 15% not docked or moored to a referent is meaningless; the counterfactual rate is needed. A SMR = 1.0 is similarly meaningless in the absence of the counterfactual rate. The indirect rate has no anchor; the SMR invites, almost forces comparison amongst hospitals.

The committee acknowledges that perhaps no one-dimensional summary can communicate the correct messages. A two-dimensional display to consider plots the (possibly stabilized) hospital rate versus the national rate for the hospital's case-mix. Points above the 45 degree line exceed the na-

tional rate, and the horizontal dimension encourages simultaneous consideration of the comparator. Two hospitals on the 45 degree line are not comparable unless they superimpose or their national rates are very close.

Alternatively, the (stabilized) SMRs can be plotted versus the national rate for the hospital's case-mix. Hospitals with points above 1.0 are in excess of their nationally computed, counterfactual rate, but two hospitals each with an SMR = 1.0 (or any other common value) in general will be associated with different national rates, reinforcing that hospitals should not be compared.

Data analysis principals suggest plots in the logarithmic scale, with axes labeled in the original scale, will be easier to read.

These types of displays may help calm the debate regarding shrinkage targets that depend on hospital volume. Since it is very likely that the counterfactual, national rates for large and small hospitals are very different, two such hospitals will have considerable X-axis separation. Yes, the low volume hospital will be shrunken more towards an SMR = 1.0 (log = 0), but such a display will dampen the urge to compare the two points.

10 Best Practices

The committee outlines best practices for a hospital profiling system starting with the 7 preferred attributes of statistical models used for publicly reported outcomes in Krumholz et al. (2006):

1. clear and explicit definition of an appropriate patient sample,
2. clinical coherence of model variables,
3. sufficiently high-quality and timely data,
4. designation of an appropriate reference time before which covariates are derived and after which outcomes are measured,
5. use of an appropriate outcome and a standardized period of outcome assessment,
6. application of an analytical approach that takes into account the multilevel organization of data,
7. disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples.

To these the committee adds,

8. data collection and reporting rules and data definitions that are actionable and minimize the opportunity for gaming,
9. high quality control and quality assurance in data collection, data definitions, analysis and reporting, (not so relevant to CMS, but very relevant in other contexts),
10. internal and external peer review of all aspects,

11. sufficient model criticism and sensitivity analyses to ensure results are sturdy with respect to reasonable departures from assumptions,
12. assessments conducted in culture and operational environment of reproducible research,
13. accurate and informative reporting,
14. periodic re-evaluation of all components,
15. to the degree possible consistent with HIPAA and other disclosure protections, all conducted in the context of reproducible research.

10.1 Reproducible Research

For scientific, workload and political reasons, it is important to put the profiling process in the context of reproducible research (see, Baggerly and Coombes, 2011; Mesirov, 2010) wherein there is an essentially seamless analytic system that starts with databases, feeds analyses that provide input to tables and graphs. In this context, all assumptions, data and analyses are completely documented and if someone wants to reproduce an analysis (possibly with some changes) they can do so without disturbing the integrity of the system. Effective reproducibility enhances credibility and transparency, thereby benefitting science, policy and communication. CMS provides the SAS code for their mortality and readmission models and thus does provide a transparent process.

11 Findings and Recommendations

The CMS charged the committee to,

- "Provide guidance on statistical approaches for accounting for clustering and variable sample sizes across hospitals when estimating hospital-specific performance metrics (e.g., mortality or readmission rates)."
- "Consider and discuss concerns commonly raised by stakeholders (hospitals, consumers, and insurers) about the use of HGLMs [Hierarchical Generalized Linear Models] in public reporting of hospital quality."

In this report, the committee has addressed issues and approaches related to the charge, and has identified and discussed additional issues with the goal of enhancing the validity and credibility of the CMS evaluations of hospital performance.

Commentary on principal criticisms of the current CMS approach

The committee has addressed the criticisms received by the CMS in response to the use of hierarchical logistic regression modeling in measure development as follows:

Criticism 1: The approach fails to reveal provider performance variation: The hierarchical modeling shrinkage effect reduces reported variation of hospital performance and renders the information not useful for consumers.

Committee view: The CMS seeks to report on systematic differences in patient outcome due to hospital quality, after removing variability in observed outcomes that is due to differences in case mix and stabilizing highly variable estimates. Even after risk adjustment for case-mix differences, inherent randomness causes directly estimated hospital effects and relative rates (that is, O/E ratios where the observed rate (O) is divided by its national model based expected rate (E)) for some hospitals to vary more than the systematic effects that are to be identified. This is especially true for hospitals with extremely low volumes, whose ratios provide little information about their underlying relative rates due to having very wide confidence intervals. Large reductions in reported variation are appropriate for hospital performance measures where the true systematic differences across all hospitals are small. The committee identifies as a top priority evaluating the option of expanding the model to include shrinkage targets that depend on hospital attributes.

Criticism 2: The approach masks performance of small hospitals: It is pointless to include small (low volume) hospitals in the calculations based on hierarchical modeling because they would get a rate close to the national mean. The hierarchical modeling methodology neutralizes small hospital performance.

Committee view: Data from small hospitals provide considerable information on associations between patient case mix and the outcome for parameter estimation in the hierarchical model; therefore, their data should be included in model building. The standard errors of hospital-specific estimates for low volume hospitals are typically large. Stabilization requires that these highly variable estimates are moved towards a model-based target to a greater degree than less variable estimates, resulting in more shrinkage for low volume hospital estimates. The overarching goal is to produce estimates that better reflect true, underlying hospital effects. As stated in the response to criticism 1, the committee identifies as a top priority evaluating the option of expanding the model to include shrinkage targets that depend on hospital attributes.

Criticism 3: The approach is based on complicated concepts and is difficult to communicate and explain to the public and to the providers. Stakeholders are familiar with the numerator and the denominator (O/E) and the output of logistic regression modeling, but the approach adopted by CMS replaces the "O" with a shrinkage estimate (referred to as "predicted" in CMS documents), and the concept is difficult to convey. In addition the concept of a hospital-specific effect is not comprehensible to most of the stakeholders.

Committee view: Some concepts and computations are more complicated than for the standard logistic regression approach, but the additional complexity allows for respecting the hierarchical structure of the data and stabilizing estimates, thereby reducing regression to the mean effects and bouncing around of provider-specific estimates. Furthermore, a fixed-effects, logistic regression model also produces hospital-specific effects. There is a continuum between the single-intercept, random effects model and the fixed-effects model with a directly

estimated intercept for each hospital, with the middle ground being occupied by a mixed effects model that includes hospital-level covariates. Therefore, barriers to comprehension of the concept are shared by all approaches. This report clarifies the principal building blocks of the approaches. The committee calls for improved communication on goals, methods, and interpretations.

Criticism 4: The evaluation of the National Quality Forum (NQF) steering committees on use of hierarchical modeling has been inconsistent, contingent on the point of view of the panelists. Therefore, it shows a lacks of consensus among statisticians and health service researchers in using hierarchical modeling for risk adjustment of outcome measures.

Committee view: The committee notes that to make progress on this issue, and possibly come to consensus, the debate must be evidence-based starting with a clear articulation of goals and ending with effective evaluation of the properties of candidate approaches. The committee recommends use of hierarchical models as an effective method to account for clustering of admissions within providers, to support valid and effective risk adjustment, and to produce stabilized estimates, although it recognizes that other approaches can accomplish these goals. This report clarifies goals and why hierarchical models are a valid approach, and focuses discussion on potential enhancements of the current CMS method.

The committee's investigation has led to the following conclusions and recommendations:

1. *Use of Hierarchical Generalized Linear Models*

   The committee concludes that Hierarchical Generalized Linear Logistic Modeling is an effective analytic approach that accounts for the structure of the data used in CMS mortality and readmission hospital metrics. The approach accommodates modeling of the association between outcomes and patient-level, pre-admission characteristics; with appropriate inclusion of hospital-level attributes, it can adjust the patient-outcome relation for potential confounding by hospital to the degree that the necessary information is available; and supports stabilizing hospital-specific performance estimates by shrinking direct estimates towards an appropriate target. The amount of shrinkage can be controlled to the extent that these controls accord with CMS' primary goal (see Recommendations 3, 4, and 5 below).

2. *Incorporation of procedure-specific volume*

   Other recommendations encourage serious consideration of including hospital-level (not procedure-specific) attributes in the national-level, case mix adjustment model and in setting shrinkage targets for stabilizing estimated hospital effects. The committee cautions that the issues related to use of procedure-specific volume are complex. Volume has a combined role as both an exogenous attribute that may be an independent predictor of quality (e.g., practice makes perfect) and an endogenous attribute that is in the causal pathway of the outcome. Furthermore, "low procedure-specific volume" may be a

marker for an inadequate risk adjustment that disadvantages hospitals with low-volume procedures.

Though evaluation of including procedure-specific volume is important, the committee recommends that higher priority be given to use of other hospital-level attributes in modeling case-mix and in producing shrinkage targets. However, regarding volume, use of procedure-specific volume from time periods prior to those used in an assessment is likely not problematic and should be explored for its ability to contribute to better-tailored shrinkage targets.

3. *Case-mix adjustment*

   (a) Patient-level attributes

      i. Consider whether the current set of patient-level attributes should be augmented, for example by including race or other demographics.

      ii. Evaluate broadening modeling approaches to include additional interaction terms among patient-level attributes.

      iii. Evaluate further broadening patient-level models through use of splines, classification and regression trees, random forests, and boosting (see section 8) to see if relative to current approaches they improve case mix adjustments by producing predictions with lower mean squared error, or improve other performance measures such as those in Efron (1978).

   It will be important to explore the extent to which alternative modeling strategies improve case-mix adjustments by producing predictions with lower mean squared error (i.e., predictions that are closer to the true structural relation), or improve other statistical attributes.

   (b) Hospital-level attributes

      The committee recommends that the CMS explore how best to include hospital attributes for two distinct purposes: 1) when developing the national-level risk model to reduce potential confounding induced by correlation between hospital and patient-level attributes; and 2) when calculating the shrinkage targets used to stabilize SMRs. Incorporating them is an accepted approach in other facility assessment settings. It is very important to note that hospital-level attributes should not set the comparator for a hospital's performance; indeed, the denominator of the SMR should depend only on a validly estimated relation between patient-level attributes and outcome. However, there may be confounding of this relationship with certain hospital characteristics, and methods to reduce this confounding should be explored.

      To reduce confounding and stabilize hospital-specific estimates, the committee proposes in appendix F.1 a Bayesian hierarchical model that adjusts for hospital-level attributes when developing the risk model, but constrains risk predictions to be for

a "typical hospital" so that hospital-level attributes play no other role in producing the expected value for a hospital. The model also allows for hospital-attribute-specific shrinkage targets to stabilize estimated SMRs.

The committee cautions that although statistical models are available to accomplish these goals, decisions as to what attributes to include and how to include them must be carefully considered. For example, covariate interactions may be needed (e.g., between hospital size and rural/urban status). Coding for candidate attributes needs to be evaluated. For example, should size be retained as a continuous attribute or be categorized? If categorized, how should the number of categories, and their cutoff values, be determined?

4. *Stabilizing estimated hospital effects*
   (a) Including hospital-level attributes in determining the shrinkage target when stabilizing estimated hospital effects is standard practice in other facility assessment settings. The committee recommends that the CMS give serious consideration to including such variables in setting shrinkage targets. To reduce potential confounding, covariate main effects and possibly interactions should be considered (e.g., shrinkage targets could be different for each of small-rural, large-rural, small-urban, and large-urban hospitals). As noted in recommendation (3b), various coding choices for candidate attributes should be explored and evaluated.
   (b) Evaluate the policy and statistical implications of replacing the single Gaussian prior distribution for the hospital-specific random effects by a more flexible class of distributions (see appendix H).
   (c) Consider supplementing posterior mean estimates with histogram estimates. These report the distribution of the SMRs with appropriate location, spread and shape (see appendix I).

5. *Readmission rates*

   Evaluate, modify and implement the method for assessing readmission rates proposed in section 7.

6. *Model assessment*

   Evaluate augmented approaches to model assessment (see section 8).

7. *Enhance reporting*

   CMS should enhance its reporting to further emphasize uncertainty and improve interpretation. The committee suggests enhancements such as: using exceedance probabilities; juxtaposing a histogram of the patient-specific risk estimates for each hospital with a histogram of the national distribution, or of the distribution for a relevant group of comparator hospitals to clarify important between-hospital differences (see section 9).

8. *Promulgate standards of conduct and communication*

   (a) Develop and communicate standards of practice for data collection, analysis and reporting for adoption by those conducting hospital comparisons.

   (b) Implement a transparent, continuous process of examining the consequences of the many, often independent, analytic choices made to ensure that what is done is as straightforward and accessible as it can be, consistent with meeting well-articulated standards for a "well-performing" quality reporting system.

9. *Transfer technology to other CMS evaluations*

   The statistical and policy issues considered in this report operate in the broad array of CMS performance measures and are relevant regardless of disease condition or process measure. Therefore, CMS should broaden its evaluations to domains other than assessment of thirty-day, post-discharge mortality and readmission rates. However, the specific choices may depend on context. For example, dialysis centers may all report a sufficient number of events so that a fixed-effects rather than a random-effects approach can be used in developing the national-level model and the SMRs.

12 Bibliography

Alexandrescu, R., Jen, M.-H., Bottle, A., Jarman, B., and Aylin, P. (2011). Logistic vs hierarchical modeling: An analysis of a statewide inpatient sample. *Journal of the American College of Surgery* **213,** 392–401.

Ash, A. and Shwartz, M. (1999). R^2: a useful measure of model performance when predicting a dichotomous outcome. *Stat Med* **18,** 375–384.

Baggerly, K. A. and Coombes, K. R. (2011). What information should be required to support clinical "omics" publications? *Clin Chem* **57,** in press. http://www.clinchem.org/cgi/doi/10.1373/clinchem.2010.158618.

Bayarri, M. J. and Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science* **22,** 322–343.

Berk, R. A. (2008). *Statistical Learning from a Statistical Perspective*. Springer-Verlag, New York.

Birkmeyer, J. D., Siewers, A. E., Finlayson, E. V. A., Stukel, T. A., Lucas, F. L., Batista, I., Welch, H. G., and Wennberg, D. E. (2002). Hospital volume and surgical mortality in the United States. *N Engl J Med* **346,** 1128–1137.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, reprinted by Springer-Verlag, 2007 edition.

Blumberg, M. S. (1987). Comments on HCFA hospital death rate statistical outliers. *Health Services Research* **21,** 715–739.

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Bunker, J. P., Forrest, W. H. J., Mosteller, F., , and Vandam, L. D., editors (1969). *The National Halothane Study: A Study of the Possible Association Between Halothane Anesthesia and Post-operative Hepatic Necrosis.* Report of the Subcommittee on Anesthesia, Division of Medical Sciences, National Academy of Sciences—National Research Council. U. S. Government Printing Office, Washington, DC.

Camilli, G., Cizek, G. J., and Lugg, C. A. (2001). Psychometric theory and the validation of performance standards: history and future perspectives. In, *Setting Performance Standards: Concepts, Methods and Perspectives., Cizek, GJ ed.* pages Mahwah, NJ: Lawrence Erlbaum Associates.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis, 3rd edition*. Chapman and Hall/CRC Press, Boca Raton, FL, 3nd edition.

CDC (2010). Your guide to the standardized infection ratio (sir). *NHSN e-news Special Edition* .

Citro, C. and Kalton, G., editors (2000). *Small-area Income and Poverty Estimates: Priorities for 2000 and beyond*. National Academy Press, Washington DC.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83,** 596–610.

Crainiceanu, C. M., Caffo, B. S., and Morris, J. (2011). *The SAGE Handbook of Multilevel Modeling*, chapter Multilevel functional data analysis. SAGE Publishing.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Adarsh, J., and Goodner, B. (2007). Spatially adaptive penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics* **16,** 265–288.

Dempster, A. P. (1988). Employment discrimination and statistical science (with discussion). *Statistical Science* **3,** 149–195.

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M., and Molyneux, D. H. (2007). Spatial modelling and the prediction of loa loa risk: decision making under uncertainty. *Annals of Tropical Medecine and Parasitology* **101,** 499–509.

Dimick, J. B., Staiger, D. O., Baser, O., and Birkmeyer, J. D. (2009). Composite measures for predicting surgical mortality in the hospital. *Health Aff (Millwood)* **28,** 1189–1198.

Dudley, R. A., Johansen, K. L., R, B., Rennie, D. J., and Milstein, A. (2000). Elective referral to high-volume hospitals: estimating potentially avoidable deaths. *Journal of the American Medical Association* **283,** 1159–1166.

Efron, B. (1978). Regression and ANOVA with Zero-One Data: Measures of Residual Variation. *Journal of the American Statistical Association* **73,** 113–121.

Ericksen, E. P. and Kadane, J. B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association* **80,** 98–131.

Fienberg, S. E. (2011). Bayesian models and methods in public policy and government settings (with discussion). *Statistical Science* **26,** 212–239.

Freedman, D. A. and Navidi, W. C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science* **1,** 3–39.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis, 2nd ed.* Chapman and Hall/CRC Press, Boca Raton, FL.

Gelman, A., van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* **61,** 74–85.

Goldman, E., Chu, P., Osmond, D., and Bindman, A. (2011). The accuracy of presnt-on-admission reporting in administrative data. *Health Services Research* **46,** Early view, on–line version.

Greiner, D. J. (2008). Causal inference in civil rights litigation. *Harvard Law Review* **122,** 533–598.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning. 2nd edition.* Springer Verlag.

Iezzoni, L. I. (1997). Assessing quality using administrative data. *Ann Intern Med* **127,** 666–674.

Iezzoni, L. I. (2003). *Risk Adjustment for Measuring Health Care Outcomes, 3rd ed.* Health Administration Press, Chicago, IL.

Jones, H. E. and Spiegelhalter, D. J. (2011). The identification of "unusual" health-care providers from a hierarchical model. *The American Statistician* **65,** 154–163.

Kipnis, P., Escobar, G. J., and Draper, D. (2010). Effect of choice of estimation method on inter-hospital mortality rate comparisons. *Medical Care* **48,** 458–465.

Krumholz, H. M., Brindis, R. G., Brush, J. E., Cohen, D. J., Epstein, A. J., Furie, K., Howard, G., Peterson, E. D., Rathore, S. S., Smith, S. C., Spertus, J. A., Wang, Y., and Normand,

S.-L. T. (2006). Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation. *Circulation* **113,** 456–462.

Krumholz, H. M., Lin, Z., Drye, E. E., Desai, M. M., Han, H. F., Rapp, M. T., Mattera, J. A., and Normand, S.-L. (2011). An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction 2011; 4: 243-252. *Circulation Cardiovascular Quality and Outcomes* **4,** 243–252.

Landrum, M., Bronskill, S., and Normand, S.-L. (2000). Analytic methods for constructing cross-sectional profiles of health care providers. *Health Services and Outcomes Research Methodology* **1,** 23–48.

Landrum, M. B. and Normand, Sharon-Lise T.and Rosenheck, R. A. (2003). Selection of related multivariate means: Monitoring psychiatric care in the Department of Veterans Affairs. *Journal of the American Statistical Association* **98,** 7–16.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2006). Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis* **1,** 915–946.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2009). Ranking of USRDS, provider-specific SMRs from 1998-2001. *Health Services and Outcomes Reseearch Methodology* **9,** 22–38.

Lin, X. (2007). Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models. *Lifetime Data Analysis* **13,** 533–544.

Lockwood, J., Louis, T. A., and McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics* **27,** 255–270.

Louis, T. A. and Zeger, S. L. (2008). Effective communication of standard errors and confidence intervals. *Biostatistics* **10,** 1–2.

Magder, L. S. and Zeger, S. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association* **91,** 1141–1151.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* **9,** 403–425.

Mesirov, J. P. (2010). Computer science. accessible reproducible research. *Science* **327,** 415–416.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68,** 179–199.

Mosteller, F. (2010). *The Pleasures of Statistics: The Autobiography of Frederick Mosteller.* Springer, New York. Edited by S. E. Fienberg and D. C. Hoaglin and J. M. Tanur.

Ni, X., Zhang, D., and Zhang, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics* **66,** 79–88.

Normand, S. L. T. and Shahian, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science* **22,** 206–226.

Paddock, S. and Louis, T. A. (2011). Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *J. Royal Statistical Society Ser. C (Applied Statistics)* **60,** DOI: 10.1111/j.1467–9876.2010.00760.x.

Paddock, S., Ridgeway, G., Lin, R., and Louis, T. A. (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational Statistics & Data Analysis* **50/11,** 3243–3262.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167,** 362–368.

Ross, J. S., Normand, S. L., Wang, Y., Ko, D. T., J., C., Drye, E. E., Keenan, P. S., Lichtman, J. H., Bueno, H., Schreiner, G. C., and Krumholz, H. M. (2010). Hospital volume and 30-day mortality for three common medical conditions. *New England Journal of Medicine* **362,** 1110–1118.

Shahian, D. M. and Normand, S.-L. T. (2003). The volume-outcome relationship: from luft to leapfrog. *Ann Thorac Surg* **75,** 1048–1058.

Shahian, D. M. and Normand, S.-L. T. (2008). Comparison of "risk-adjusted" hospital outcomes. *Circulation* **117,** 1955–1963.

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage, hierarchical models. *Journal of the Royal Statistical Society, Series B* **60,** 455–471.

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 Serv Res* **45,** 1148–1167.

Spencer, G., Wang, J., Donovan, L., and Tu, J. V. (2008). Report on Coronary Artery Bypass Surgery in Ontario, Fiscal Years 2005/06 and 2006/07. Technical report, Institute for Clinical Evaluative Sciences.

Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *J. Royal Statistical Society, Ser. B* **64,** 583–639.

Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., and Grigg, O. (2012). Statistical methods for healtcare regulation: rating, screening and surveillance. *J. Roy. Statist. Soc. Ser. A* **175,** 1–25.

Tomberlin, T. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association* **83,** 309–321.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Chapman and Hall, CRC Press.

Whoriskey, P. (2011). Florida to link teacher pay to students' test scores. *Washington Post, March 22, 2011* .

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Appendices

A Glossary of Acronyms

AIC: Akaiki's information criterion

AMI: acute myocardial infarction

AUC: area under the curve

BIC: Bayesian information criterion

CABG: coronary artery bypass grafting

CCN: [Ontario] Cardiac Care Network

CHI: Commission for Health Improvement [UK]

CMS: Centers for Medicare and Medicaid Services

COPD: chronic obstructive pulmonary disease

COPSS: Committee of Presidents of Statistical Societies

DFFITS: difference in fits (when a case is included versus omitted from the fit of the model)

DIC: deviance information criterion

DRG: diagnostic related group

EDF: empirical distribution function

GLMs: Generalized Linear Models

HF: heart failure

HGLMs: Hierarchical Generalized Linear Models

HIPAA: Health Insurance Portability and Accountability Act

HWR: Hospital-Wide Readmission

ICC: Intra-Class Correlation

ICU: intensive care unit

LOS: length of stay

MI: myocardial infarction

MLE: maximum likelihood estimator

NQF: National Quality Forum

PCI: percutaneous coronary interventions

PRESS: prediction sum of squares

QIR: Quality Improvement Program

RSMRs: risk-standardized mortality rates

RSRRs: risk-standardized readmission rates

SD: standard deviation

SE: standard error

SMR: Standardized Mortality Ratio

TEP: Technical Expert Panel

YNHHSC/ CORE: Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation

B Committee Member Biographical Sketches

ARLENE S. ASH, PHD in Mathematics, is a methods expert on risk adjustment and value assessment in health. She is Professor and Division Chief of Biostatistics and Health Services Research in the Department of Quantitative Health Sciences at the University of Massachusetts Medical School in Worcester, MA. She is a Fellow of AcademyHealth and of the American Statistical Association (ASA), and recipient of the ASA Health Policy and Statistics Section's inaugural lifetime achievement award. In 1988 she received the Administrator's Citation for her work on Medicare capitation issues. In 1996 she helped found DxCG, Inc. (now, the science division of Verisk Health). Her role in developing tools and fostering the world-wide dissemination of "Risk-Based Predictive Modeling" in health was recognized in AcademyHealth's 2008 Health Services Research Impact Award. Her over 150 published papers encompass a wide range of interests, with special attention to gender and racial/ethnic disparities in health care delivery and outcomes, how payment systems affect care, and how providers can be monitored to detect quality. Her current research focuses on calculating payments that provide patient-centered medical homes with global budgets tailored to the health of their patient panels and methods for assessing the extent to which practices perform better-than-expected with respect to various metrics.

STEPHEN E. FIENBERG, PHD is Maurice Falk University Professor of Statistics and Social Science at Carnegie Mellon University, with appointments in the Department of Statistics, the Machine Learning Department, the Heinz College, and Cylab. His principal research interests lie in the development of statistical methodology, especially for problems involving categorical variables. He has worked on the general statistical theory of log-linear models for categorical data, including approaches appropriate for disclosure, limitation estimating the size of populations, network analysis, and Bayesian approaches to the analysis of contingency tables and network structures. He has written or edited over 20 books and 400 scientific publications. Fienberg has served as a Vice-President of the American Statistical Association and as President of the Institute of Mathematical Statistics and the International Society for Bayesian Analysis. He is the Editor-in-chief of the Journal of Privacy and Confidentiality, and serves on the editorial board of the Proceedings of the National Academy of Sciences and currently serves as Co-Chair of the NAS-NRC the Report Review Committee, 2008–2012. He is an elected member of the National Academy of Sciences, and a fellow of the American Academy of Arts and Sciences and the Royal Society of Canada.

THOMAS A. LOUIS, PHD is Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health. Research includes Bayesian methods, risk assessment; environmental and public policy, health services, genomics, clinical trials and malaria control. He has published over 280 articles, books/chapters, monographs and discussions. Professor Louis is an elected member of the International Statistical Institute, a Fellow of the American Statistical Association and of the American Association for the Advancement of Science. From 2000 through 2003, he was coordinating editor of The Journal of the American Statistical Association (JASA) and is currently a co-editor of Biometrics. He has served as president of the Eastern North American Region of the International

Biometric Society (IBS) and President of the IBS. He has chaired the ASA section of Bayesian Statistical Science and currently is chair of the American Association for the Advancement of Science Statistics section. He is a member of the Board of Scientific Counselors, NIH/NIEHS. National Academy panel and committee service includes the Committee on National Statistics, the Committee on Applied and Theoretical Statistics, the Panel on Estimates of Poverty for Small Geographic Areas, the Panel on Formula Allocation of Federal and State Program Funds (chair), the Board of the Institute of Medicine's Medical Follow-up Agency, the IOM Panel to Assess the Health Consequences of Service in the Persian Gulf War, the Committee on the use of Third Party Toxicity Research.

SHARON-LISE T. NORMAND, PHD is Professor of Health Care Policy (Biostatistics) in the Department of Health Care Policy at Harvard Medical School and Professor in the Department of Biostatistics at the Harvard School of Public Health. Her research focuses on the development of statistical methods for health services and outcomes research, primarily using Bayesian approaches, including causal inference, provider profiling, item response theory analyses, latent variables analyses, multiple informants analyses, and evaluation of medical devices in randomized and non-randomized settings. She serves on several task forces for the American Heart Association and the American College of Cardiology, was a consultant to the US Food and Drug Administration's Circulatory System Devices Advisory Panel after serving a four-year term on the panel, is a member of the Medicare Evidence Development and Coverage Advisory Committee, a consultant to the Centers for Medicare and Medicaid Services on the development of their hospital outcome models, and is Director of Mass-DAC, a data coordinating center that monitors the quality of all adult cardiac surgeries and coronary interventions in all Massachusetts acute care hospitals. Dr. Normand has served on several editorial boards including Biometrics, Statistics in Medicine, Health Services and Outcomes Research Methodology, Psychiatric Services, and Cardiovascular Quality and Outcomes. She was the 2010 President of the Eastern North American Region of the International Biometrics Society and is Vice Chair of the Patient Centered Outcomes Research Institute's Methodology Committee. Dr. Normand earned her Ph.D. in Biostatistics from the University of Toronto, holds a Master's of Science as well as a Bachelor of Science degree in Statistics, and completed a post-doctoral fellowship in Health Care Policy at Harvard Medical School. She is a Fellow of the American Statistical Association, a Fellow of the American College of Cardiology, a Fellow of the American Heart Association, and an Associate of the Society of Thoracic Surgeons.

THÉRÈSE A. STUKEL, PHD is a Senior Scientist at the Institute for Clinical Evaluative Sciences (ICES), Toronto; Professor of Biostatistics and Health Services Research, The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth Medical School, Hanover NH; Professor of Health Policy, Management and Evaluation, University of Toronto. She was statistical director of the Dartmouth Atlas of Health Care from 1995 to 2003 and co-authored two influential publications on the U.S. healthcare system demonstrating that higher healthcare spending did not lead to better outcomes. Her current research interests are in the effects of health system resources and

organization on delivery of care and outcomes in Canada and the U.S., particularly whether higher health care spending is associated with better health systems outcomes. She has also focused on methods for the analysis of observational studies, particularly the use of instrumental variables to remove unmeasured confounding and survival bias. With the support of a CIHR Team grant, she is creating virtual physician networks in Ontario and evaluating their efficiency (quality vs costs) in managing patients with chronic disease. She has published over 160 peer-reviewed articles in medical and statistical journals. She was nominated Fellow of the American Statistical Association in 2007.

JESSICA UTTS, PHD is Professor and Chair of the Department of Statistics at the University of California, Irvine. Her research interests include the application of statistical methods to medicine and psychology, including both Bayesian and non-Bayesian methods. She is also a strong advocate for the promotion of statistical literacy, especially for decision-makers, and has authored or co-authored three textbooks with a focus on statistical literacy as well as several journal articles. She is a Fellow of the American Statistical Association, the American Association for the Advancement of Science, the Institute of Mathematical Statistics and the Association for Psychological Science, and an elected member of the International Statistical Institute. She has served as President of the Western North American Region of the International Biometric Society, as Chair of the Committee of Presidents of Statistical Societies (COPSS) and as Chair of the American Statistical Association Section on Statistical Education. She is currently a Vice President of the International Association for Statistical Education, a member of the Board of Directors of the American Statistical Association and of the National Institute of Statistical Sciences, and Chair of the Board of the Consortium for the Advancement of Undergraduate Statistics Education.

C CMS Statement of Work and Committee Interpretation

The Committee of Presidents of Statistical Societies, with funding from the Centers for Medicare and Medicaid Services (CMS), through a subcontract with YNHHSC/CORE, was guided by the following statement of work and subsequent committee interpretation.

C.1 Statement of work

The Centers for Medicare and Medicaid Services (CMS) has contracted with YNHHSC/CORE to develop a white paper on the appropriateness of statistical modeling for hospital outcomes measures. Specifically, CMS has indicated the paper should address the statistical modeling used in current CMS publicly reported outcomes measures for mortality and readmission (hierarchical generalized linear models [HGLMs]) and the concerns about this statistical approach that are frequently cited by stakeholders (hospitals, consumers, quality measure developers, and insurers). To develop this paper, YNHHSC/CORE intends to subcontract with a group of individuals recognized by professional statistical societies to provide guidance on statistical approaches for assessing hospital quality based on patient outcomes. Accordingly, the scope of work for this subcontract is to:

1. Provide guidance on statistical approaches for accounting for clustering and variable sample sizes across hospitals when estimating hospital-specific performance metrics (e.g., mortality or readmission rates);

2. Consider and discuss concerns commonly raised by stakeholders (hospitals, consumers, and insurers) about the use of HGLMs in public reporting of hospital quality;

3. Develop a draft white paper regarding these findings for broad dissemination;

4. Solicit comments from the statistical community; and

5. Prepare a final white paper for dissemination.

The final paper should be completed no later than Sept. 15, 2011.

C.2 Committee interpretation

The committee has reviewed the Statement of Work from CMS and provide the following interpretation of it as a clarification of our plans.

With funding from The Centers for Medicare and Medicaid Services (CMS) via a sub-contract with YNHHSC/CORE, a group of individuals recognized by professional statistical societies will develop a white paper on the appropriateness of statistical modeling for measures of hospital outcomes. The committee shall address the statistical modeling used in current CMS publicly reported outcomes measures for mortality and readmission (hierarchical generalized linear models [HGLMs]), and will take into account the concerns about this statistical approach as formally communicated by stakeholders (hospitals, consumers, quality measure developers, and insurers). The white paper will not consider issues related to developing appropriate case-mix adjustments and so will focus on use of risk-adjusted values.

Accordingly, by September 15, 2011 the committee shall:

1. Provide guidance on statistical approaches for accounting for clustering and variable sample sizes across hospitals when estimating hospital-specific performance metrics (e.g., mortality or readmission rates) or ranks;

2. This guidance will take into account concerns formally communicated by stakeholders (hospitals, consumers, and insurers) about the use of HGLMs in public reporting of hospital quality;

3. Develop a draft white paper regarding these findings for broad dissemination;

4. Conduct a timely and independent review of the draft document

5. Prepare the final report for dissemination.

D Background provided by the CMS

(This information was provided by Lein Han, PhD)

Since 2008 CMS has been publicly reporting condition-specific outcome measures such as risk adjusted all-cause mortality and readmission measures for AMI, HF and Pneumonia for the CMS Inpatient Quality Reporting (IQR) program. These measures are developed by CMS and endorsed by the National Quality Forum (NQF). Currently there is an increasing need for CMS to develop more outcome measures. Several sections of the Affordable Care Act of 2010 require the Secretary to develop and implement outcome measures to meet a variety of program needs. Concurrent with the CMS's effort to expand its outcome measurement development work to meet the congressional mandates, CMS sought to convene a Technical Expert Panel (TEP) meeting on Statistical Modeling for Outcome Measures. The TEP will serve as the mechanism deployed by CMS to obtain a consensus on the appropriate statistical modeling methodology for outcome measure development. The goals of the statistical TEP meeting are to inform CMS outcome measurement development work. CMS expects a White Paper from the TEP that would recommend to CMS the most appropriate statistical modeling approach to the development of hospital-specific risk-adjusted outcome measures for the CMS public reporting and value-based pursing initiatives in order to (1) meet the congressional requirements and (2) address concerns stakeholders have expressed regarding the use of hierarchical logistic modeling.

CMS uses hierarchical modeling to produce risk-adjusted hospital-specific measures for mortality, readmission, and complications. CMS has received some push-back from the stakeholders as well as research communities, such as statisticians and health service researchers, regarding use of this methodology. Because the TEP includes representatives from various prominent statistical societies in the nation, CMS expects that the statistical societies would clarify their position and provide guidance on the appropriate statistical modeling for risk adjusting outcome measures for public reporting and VBP at the provider level. Providers include hospitals, physicians, or health plan. The purpose is to assist in standardizing statistical modeling for outcome measure development across measure developers in the public and private sectors, and hence providing consistent and comparable information for (1) providers, (2) consumers, and (3) government to pay for performance.

The following describes briefly the congressional mandates for outcome measure development and the criticisms that CMS received regarding use of hierarchical modeling. For details of the mandates, please review the Affordable Care Act. The information intends to provide a background/context for the TEP's deliberation.

D.1 Affordable Care Act

Section 3025 establishes the Hospital Readmission Reduction Program and requires that the Secretary use the CMS NQF-endorsed readmission measures for high-cost, high volume conditions to be selected by the Secretary. Through the FY 2012 IPPS/LTCH PPS final rule, the Secretary

proposed and finalized the conditions acute myocardial infarction (AMI), heat failure (HF), and Pneumonia for use in the Hospital Readmission Reduction Program. section 3025 further provides, to the extent practicable, for the development of additional readmission measures for conditions identified by the Medicare Payment Advisory Commission in its June 2007 report to Congress for calculating the Hospital Excess Readmission Ratio as part of the basis for the hospital diagnostic related group (DRG) payment.

Section 3025, and Section 399KK establish the Quality Improvement Program (QIP) It requires that the Secretary identify hospitals with high risk-adjusted readmission rates and make available to them a program to reduce readmissions and improve patient safety through the use of patient safety organizations. section 399KK further requires that the Secretary use the readmission measures selected for section 3025 to identify hospitals with high risk-adjusted readmission rates.

In addition to the above sections, CMS plans to make use of condition-specific readmission measures to support a number of quality improvement and reporting initiatives. CMS is also in the process of developing a Hospital-Wide Readmission (HWR) measure that CMS may consider for implementation in quality programs through future rulemaking. Below are several examples of the criticisms received in response to the use of hierarchical logistic regression modeling in measure development by CMS.

1. Fails to reveal provider performance variation: The shrinkage effect of the hierarchical modeling *reduces* the variation of the hospital performance and renders the information not useful for the consumers

2. Masks the performance of small hospitals: It is pointless to include the small hospitals in the calculation based on hierarchical modeling because the small hospitals would get a rate close to the national mean. The methodology *neutralizes* small hospital performance.

3. A difficult concept to communicate or explain to the public and the providers: stakeholders are familiar with the result (the numerator and the denominator) of the logistic regression modeling, commonly referred to as *the Observed over the Expected* (O to E). Because the approach adopted by CMS replaces the O with a shrinkage estimate (referred to as *predicted* in CMS documents), the concept is difficult to convey to the public and the stakeholders. In addition the concept of *hospital-specific* effect is not comprehensible to most of the stakeholders

4. The evaluation of the NQF steering committees on the use of hierarchical modeling has been inconsistent contingent on the point of view of the panelists. Therefore it shows a lacks of consensus in using hierarchical modeling for risk adjustment for outcome measures among statisticians and health service researchers.

E The CMS Approach and Extensions

The statistical model adopted by CMS uses variance components to characterize the potential sources of variation in the outcome at the patient level, after accounting for patient-level risk factors and hospital-level variation. To permit both types of variation, CMS uses a hierarchical model in which the first stage specifies a probability distribution for the risk of the outcome for patients treated within the same hospital. The log-odds of the event for a patient treated within a hospital is a function of patient-specific admission characteristics and a hospital-specific intercept. In the second stage, the hospital-specific intercepts arise from a normal distribution with a common mean and variance component. This modeling strategy is particularly useful for two reasons. First, it accommodates a range of plausible alternatives. If the hospital variance component is zero, then this implies that all differences in outcomes are explained entirely by patient-level differences in risk factors and sampling variability. If the between-hospital variance component is very large, then this implies that there are large differences in outcomes across hospitals, possibly explainable by other factors. However, in the absence of a model that provides at least a partial explanation, the hospital effects are completely unrelated and so data from each hospital should be modeled separately. Finally, a hospital variance component between these two extremes implies moderate differences in outcomes across hospitals and possibly some relations among the hospital effects. The second key reason why the modeling strategy is useful relates to the multiplicity problem. CMS is interested in making inferences on many parameters, e.g., on the order of 2000 to 3000 hospital-specific ones. If the model assumptions hold, then the many-parameter problem reduces to a two parameter problem, involving the common mean and the between-hospital variance component.

The CMS hospital performance measure is a risk standardized *rate* using the population of patients treated at the hospital. The performance measure uses parameters from the hierarchical model to determine a numerator and a denominator. The numerator reflects what the outcomes were at each hospital, but rather than using the observed number of events at the hospital, the individual risk probabilities for each patient in the hospital are computed by multiplying the risk coefficients by the patients' risk factors, adding the stabilized, hospital specific intercept, and then summing. This yields the (conditional) expected total number of events for that hospital. CMS uses the conditional expected total number of events rather than the observed number events to avoid regression to the mean and to gain precision for lower volume hospitals. The denominator reflects what the outcome *would have been* at a hospital given its actual distribution of patients but replacing the observed outcomes with those estimated from all hospitals in the sample. The denominator sums the individual risk probabilities for each patient within a given hospital, using the risk coefficients estimated from the regression model, the patients' distributions of risk factors, and the overall intercept. This yields the expected total number of events for that hospital. The indirectly standardized ratio, the numerator divided by the denominator, represents the outcome for a hospital's specific distribution of patients had those patients been treated by an average provider. The ratio is converted to a rate by multiplying the ratio by the national percent experiencing the event.

E.1 Current CMS approach: Technical details

The CMS approach involves determination of a hospital-specific estimator rather than hypothesis *testing*. Let Y_{ij} denote a binary outcome for the jth patient treated at the ith hospital, $\mathbf{x}_{ij}$ a vector of patient-specific characteristics, and n_i the number of cases treated at hospital i for $i = 1, 2, \cdots, I$ hospitals. Assuming $Y_{ij} = 1$ for the jth patient treated in the i^{th} hospital and 0 otherwise, the CMS model assumes the following:

$$\begin{aligned} [Y_{ij} \mid \beta_{0i}, \alpha, \mathbf{x_{ij}}] &\overset{ind}{\sim} \text{Bern}(p_{ij}) \text{ where } \text{logit}(p_{ij}) = \beta_{0i} + \boldsymbol{\alpha}\mathbf{x}_{ij} \\ [\beta_{0i} \mid \mu, \tau^2] &\overset{iid}{\sim} N(\mu, \tau^2). \end{aligned} \tag{2}$$

and with $\mathbf{x}_{ij}$ denoting a vector of patient-specific admission characteristics. To increase sample sizes, CMS uses a 3-year observational period of data for each hospital, denoted n_i. In equation (3) τ^2 represents between-hospital variation after accounting for what the patient looked like at admission. Through a probability model, the CMS approach permits underlying hospital quality to vary around an overall mean effect denoted by μ. If there are no between-hospital differences in the outcome beyond that captured by the $\mathbf{x}_{ij}$, then $\tau^2 = 0$ and $\beta_{01} = \beta_{02} = \cdots = \beta_{0I} = \mu$. In this case, any observed differences in the unadjusted hospital outcomes would be due to case-mix (patient factors). While it is almost certain that $\tau^2 > 0$, a question is whether τ is small enough to ignore.

An implicit assumption in the model defined by equation (2) is that hospital mortality is independent of the number of patients treated at the hospital, after conditioning on patient characteristics. The hospital-specific estimator uses as its counter-factual population subjects with the same case-mix as those observed at the hospital, e.g., $\mathbf{x}_{ij}$, and *risk effects* quantified by the national average, e.g., μ and $\boldsymbol{\alpha}$.

$$\begin{aligned} \theta(\mathbf{x})_i &= \frac{\sum_{j=1}^{n_i} E(Y_{ij} \mid \beta_{0i}; \mathbf{x}_{ij}, \mu, \alpha, \tau^2)}{\sum_{j=1}^{n_i} E(Y_{ij} \mid \mathbf{x}_{ij}, \mu, \alpha, \tau^2)} \times \bar{Y} \\ &= SMR_i \times \bar{Y} \end{aligned} \tag{3}$$

where

$$\bar{Y} = \frac{\sum_{i,j} Y_{ij}}{\sum_i n_i}.$$

The expectation in the numerator of equation (3) integrates over the posterior distribution of β_{0i} using the model in equation (2). The expectation in the denominator integrates over the prior distribution. In practice (μ, τ, α) are replaced by their estimated values.

An important characteristic of the CMS hospital-specific estimator involves its *comparator*; each hospital performance is compared to a population having the *same* case mix as itself. This feature protects against extrapolation outside of the hospital's treated case-mix; importantly, pair-wise comparisons of $\theta(\mathbf{x})_i$ with $\theta(\mathbf{x})_j$ would only be meaningful to the extent that the distributions of $\mathbf{x}_i$ and $\mathbf{x}_j$ overlap and are balanced.

E.1.1 Estimation

The CMS model uses the *SAS* procedure *PROC GLIMMIX* to estimate the hierarchical model parameters and the bootstrap to calculate estimates for $\theta(\mathbf{x})_i$ as well as 95% interval estimates (Ross et al., 2010). CMS makes the SAS code publicly available.

Table 1: **CMS Bootstrap Algorithm.** Procedures used to estimate the CMS hospital-specific estimates of quality. b indexes the bootstrap sample.

1) Sample I hospitals with replacement.
2) Fit model in Equations (2) - (3) using all cases within each sampled hospital. Each hospital is treated as distinct. Using *Glimmix*, calculate (a) The hospital fixed effects: $\hat{\alpha}^b$. (b) The parameters governing the hospital-specific random effects distribution, $\hat{\mu}^b$, and $\hat{\tau}^{2(b)}$. (c) The set of hospital-specific estimates andh corresponding variances, $\{\hat{\beta}_{0i}, \widehat{\text{var}}(\beta_{0i})$; $i = 1, 2, \cdots, I\}$. If a hospital is sampled more than once, randomly select one set of hospital-specific estimates and hyper-parameters.
3) Simulate a hospital random effect by sampling from the posterior distribution of the hospital-specific distribution obtained in Step 2(c). The posterior is approximated by a normal distribution, $\hat{\beta}_{0i}^{b*} \sim N(\hat{\beta}_{0i}, \widehat{\text{var}}(\beta_{0i}))$ for the unique set of hospitals obtained from Step 1.
4) Within each unique hospital i sampled in Step 1, and for each observation j in that hospital, calculate $\hat{\theta}(\mathbf{x})_{\mathbf{i}}^{(\mathbf{b})}$ using $\hat{\alpha}^{b(k)}$ from Step 2 and using $\hat{\beta}_{0i}^{b*}$ from Step 3, $\hat{\theta}^b(\mathbf{x})_i = \log \left[\frac{\sum_{j=1}^{n_i} \text{logit}^{-1}(\hat{\beta}_{0i}^{b*(x)} + \hat{\alpha}^{b(x)}\mathbf{x}_{ij})}{\sum_{j=1}^{n_i} \text{logit}^{-1}(\hat{\mu}^{b(x)} + \hat{\alpha}^{b(x)}\mathbf{x}_{ij})} \right] + \log(\bar{Y})$ and for $i = 1, 2, \cdots, I$. Logarithms are taken to ensure positivity of the estimates.

F Modeling to Allow for Hospital-Attribute Shrinkage Targets

In order to allow hospital-level attributes to influence the shrinkage targets for stabilization of the SMR in the numerator model, there must be some de-linkage of the risk adjustment and the stabilization components. The following two approaches are strategically different. Both require careful implementation and comparison to the current model when there are no hospital-level covariates included. Comparisons should be empirical, involving the same datasets, and property-based, involving simulations.

F.1 A generalization of the current CMS model

The committee recommends that hospital-level attributes be used in developing the national-level risk model to reduce potential confounding induced by correlation between hospital and patient-level attributes, and that they also be used to develop shrinkage targets when stabilizing SMRs. Regarding the former role, it is absolutely the case that hospital-level attributes should not set the comparator for a hospital's performance; that should depend only on a validly estimated relation between patient-level attributes and outcome. However, there may well be confounding of this relation by hospital, and it is important to have a way to reduce this confounding.

To address the confounding reduction and stabilization goals, the committee proposes the following Bayesian hierarchical model that adjusts for hospital-level attributes when developing the risk model, but risk predictions for individual patients to estimate the expected rates are constrained to be for a "typical" hospital. Hospital-level attributes play no other role in producing the expected value for a hospital. The model also allows for hospital-attribute specific shrinkage targets to stabilize estimated SMRs. The committee cautions that though statistical models are available to accomplish these goals, the decision on what attributes to include needs to be carefully considered. These issues are discussed in other sections of this report.

The following development follows directly from that for the current CMS model in appendix E.1. Let $Y_{ij} = 1$ for the jth patient treated in the i^{th} hospital and 0 otherwise. Assume the following,

$$\begin{aligned} [Y_{ij} \mid \beta_{0i}, \alpha, \mathbf{x_{ij}}] &\overset{ind}{\sim} \text{Bern}(p_{ij}) \text{ where } \text{logit}(p_{ij}) = \beta_{0i} + \boldsymbol{\alpha}\mathbf{x}_{ij} \\ [\beta_{0i} \mid \mu, \tau^2, \gamma, \mathbf{z_i}] &\overset{iid}{\sim} N(\mu + \gamma\mathbf{z_i}, \tau^{\mathbf{2}}), \end{aligned} \tag{4}$$

with $\mathbf{x}_{ij}$ denoting a vector of patient-specific, admission characteristics and $\mathbf{z_i}$ denoting a vector of hospital-level attributes that are to be used to develop shrinkage targets for the numerator of the SMR. Note that, though the $\mathbf{z_i}$ do appear in the prior distribution for the β_{0i}, as clarified below they are not used to adjust the reference population when computing an SMR.

The proposed, hospital-specific estimator is,

$$\begin{aligned}\theta(\mathbf{x})_i &= \frac{\sum_{j=1}^{n_i} E(Y_{ij} \mid \beta_{\mathbf{0i}}; \mathbf{x_{ij}}, \mathbf{z_i}, \mu, \tau^{\mathbf{2}}, \alpha, \gamma)}{\sum_{j=1}^{n_i} E(Y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}^*, \mu, \alpha, \tau^{\mathbf{2}}, \gamma)} \times \bar{Y}, \qquad (5)\\ &= SMR_i \times \bar{Y}\end{aligned}$$

where

$$\bar{Y} = \frac{\sum_{i,j} Y_{ij}}{\sum_i n_i}$$

and $\mathbf{z}^*$ is chosen to satisfy

$$\bar{Y} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} E(Y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}^*, \mu, \alpha, \tau^{\mathbf{2}}, \gamma). \qquad (6)$$

The expectation in the numerator of equation (5) integrates over the posterior distribution of β_{0i} using the model in equation (4). The expectation in the denominator of equation (5) and in equation (6) integrates over the prior distribution with a fixed set of hospital-level attributes; $\mathbf{z_i} \equiv \mathbf{z}^*$ If $\mathbf{z}$ has more than one component, the there will be a variety of $\mathbf{z}^*$ that satisfy equation (6), but they will all produce the same predicted number of events for a hospital. In practice $(\mu, \tau, \alpha, \gamma)$ are replaced by their estimated values. A simple adaptation of the methods used for the current CMS model should work for this expanded model.

The current CMS model corresponds to the exclusion of hospital-level covariates. While hospital-level attributes enter into the prior for the β_{0i}, they do not change the referent population. An SMR = 1.0 is still interpreted as the hospital is performing at the nationally predicted level for a counterfactual hospital with the same case-mix. Inclusion of the $\mathbf{z_i}$ increases the degree of adjustment for hospital confounding over that provided by the random effects specification for the β_{0i} and so the estimated values for (μ, τ^2, α) will be different from those for the current model, but the estimation goal remains unchanged. Inclusion of the hospital-level attributes in the risk model will produce an estimated τ^2 that is smaller than that produced by the current CMS model. Thus, the weight on the direct estimate will be smaller, but the shrinkage will be towards the target determined by hospital-level attributes rather than to an overall, hospital-independent value.

F.2 A two-stage approach

The following, two-stage approach is included for illustrative purposes. It clarifies the separate roles of the risk prediction as stabilization components of the assessment process. The first stage extracts estimated hospital effects (denoted by $\hat{\beta}_{0i}$) and their standard errors. These are then analyzed by a random effects model.[1]

[1] Before statistical theory and computing to estimate unified models were available, this approach was in common use, sometimes referred to as "the NIH model."

F.2.1 The stage 1, direct SMR model

Assume that the CMS national-level hierarchical logistic regression (see appendix E.1) is available to produce the national-level expected deaths. Specifically, denote the denominator of equation (3) by $\mathrm{ED}_i(\mathbf{x_{ij}}, \mu, \alpha, \tau^{\mathbf{2}})$ and using the observed number of hospital-specific deaths (Y_{i+}) as the numerator produce the directly estimated SMR (denoted dSMR),

$$dSMR_i = \frac{Y_{i+}}{\mathrm{ED}_i(\mathbf{x}_{ij}, \mu, \alpha, \tau^2)} \ , i = 1, \ldots, I. \tag{7}$$

To estimate the hospital effect, find the value of μ in equation (7) (denoted by $\hat{\beta}_{0i}$) so that,

$$\mathrm{ED}_i(\mathbf{x}_{ij}, \hat{\beta}_{0i}, \alpha, \tau^2) = Y_{i+} \ .$$

So, we have the relations,

$$dSMR_i = \frac{Y_{i+}}{\mathrm{ED}_i(\mathbf{x}_i, \mu, \alpha)} = \frac{\mathrm{ED}_i(\mathbf{x}_{ij}, \hat{\beta}_{0i}, \alpha, \tau^2)}{\mathrm{ED}_i(\mathbf{x}_{ij}, \mu, \alpha, \tau^2)} \ . \tag{8}$$

The $(\hat{\beta}_{0i} - \mu)$ are *hospital effects* (in fact, estimated hospital *fixed effects*). They consolidate over the patient mix all hospital influences that operate after having adjusted for case mix. Their values directly translate into dSMR values; for example if $\hat{\beta}_{0i} = \mu$, then $dSMR_i = 1.0$, if $\hat{\beta}_{0i} < \mu$, then $dSMR_i < 1.0$, etc. With $I > 1800$, there are a large number of them; some quite stably estimated, others quite unstable.

To compute the full sampling distribution for the $\hat{\beta}_{0i}$ one could use a simulation that repeatedly generates Y_{+i} from the national-level risk model (details omitted), and then use it to estimate the the standard errors of the $\hat{\beta}_{0i}$ (denoted by σ_i) and confidence intervals for them and thereby for the dSMRs. For example, a 95% interval corresponds to replacing $\hat{\beta}_{0i}$ in equation (8) by $\hat{\beta}_{0i} \pm 1.96 \times \hat{\sigma}_i$.

F.2.2 The stage 2, (empirical) Bayes model

As in appendix F.1, let $\mathbf{z_i}$ be a vector of hospital-level attributes. Using the normal-distribution model (see appendix H for other models) write,

$$\begin{aligned} [\beta_{0i} \mid \mu, \tau^2, \gamma, \mathbf{z_i}] &\sim N(\mu + \gamma \mathbf{z_i}, \tau^{\mathbf{2}}) \\ [\hat{\beta}_{0i} \mid \beta_{0i}, \sigma_i^2] &\sim N(\beta_{0i}, \sigma_i^2) \end{aligned} \tag{9}$$

In the empirical Bayes approach all hospitals contribute information for estimating the prior mean, variance and regression slopes (the γ). Specifically, the collection of $(\hat{\beta}_{0i}, \hat{\sigma}_i)$ are used to produce $(\hat{\mu}, \hat{\tau}^2, \hat{\gamma})$. The estimated prior variance $(\hat{\tau}^2)$ quantifies the amount of variation in the $\hat{\beta}_{0i}$ that is not explained by the $\hat{\sigma}_i^2$ or the $\gamma \mathbf{z_i}$. Recursive estimation is needed (See Carlin and Louis, 2009, for details). Though more weight is given to the low variance $\hat{\beta}_{0i}$ than to the highly variable ones, the weights are flatter than those produced by $1/\hat{\sigma}_i^2$ and even low volume (large $\hat{\sigma}_i^2$) hospitals contribute substantially to these estimates.

The "plug-in" posterior means are the shrunken estimates,

$$\hat{\beta}_{0i}^{pm} \;=\; E(\beta_{0i} \mid \hat{\beta}_{0i}, \hat{\sigma}_i^2, \hat{\mu}, \hat{\tau}^2, \hat{\gamma}, \mathbf{z_i}) = (\hat{\mu} + \hat{\gamma}\mathbf{z_i}) + \frac{\hat{\tau}^{\mathbf{2}}}{\hat{\tau}^{\mathbf{2}} + \hat{\sigma}_{\mathbf{i}}^{\mathbf{2}}} \{\hat{\beta}_{\mathbf{0i}} - (\hat{\mu} + \hat{\gamma}\mathbf{z_i})\}$$

and the plug-in posterior variances are,

$$V(\beta_{0i} \mid \hat{\beta}_{0i}, \hat{\sigma}_i^2, \hat{\mu}, \hat{\tau}^2, \hat{\gamma}, \mathbf{z_i}) \;=\; \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_i^2} \;\hat{\sigma}_i^2 \;.$$

The amount of shrinkage depends on the relation between $\hat{\sigma}_i^2$ and $\hat{\tau}^2$. Highly variable direct estimates are given less weight with more allocated to the hospital-attribute specific shrinkage target $(\hat{\mu} + \hat{\gamma}\mathbf{z_i})$. Because $\hat{\tau}^2$ measures unexplained variation at the hospital level around the hospital-attribute specific mean, hierarchical models *allow for* between-hospital variation, but do not require that it exists. For example, if $\hat{\tau}^2$ were forced to equal 0, $\hat{\beta}_{0i}^{pm} \equiv (\hat{\mu} + \hat{\gamma}\mathbf{z_i})$; if $\hat{\tau}^2$ were forced to be, say, 1000 times larger than the largest $\hat{\sigma}_i^2$, $\hat{\beta}_{0i}^{pm} \doteq \hat{\beta}_{0i}$. This latter relation with the posterior mean equal to the direct estimate and the posterior variance equal to the sampling variance produces the directly estimated SMRs (the dSMRs) and has been termed "frequentist Bayes." It is proposed as a reporting option in sections 7 and 9.

If no hospital-level attributes are included in the prior ($\gamma \equiv 0$), then the estimates for hospitals with large $\hat{\sigma}_i^2$ are moved close to the national mean ($\hat{\mu}$), producing estimated SMRs close to 1.0. This feature generates a principal point of debate regarding whether and how to use hierarchical models with associated shrinkage.

G The Michigan, Canadian and United Kingdom Approaches

The subsequent subsections outline the Michigan, Canadian and United Kingdom approaches to hospital profiling.

G.1 The Michigan Approach

Since 1995, the University of Michigan's Kidney Epidemiology and Cost Center has produced dialysis facility-specific information (www.dialysisreports.org) for CMS. CMS displays this information on their Dialysis Facility Compare website, similar to their Hospital Compare website. In addition to process measures, each facility's actual patient survival is compared to its expected patient survival. Data from each facility cover a three-year observation window. Covariates in this model include a patient's age, race, sex, diabetes, and years on dialysis, whether they had other health problems when they started dialysis, additional diagnoses such as cancer or heart problems and body size. Facilities' survival rates are categorized as better than expected, expected, or worse than expected.

G.2 The Canadian Approach

Canadian public hospital reporting is limited to coronary artery bypass graft (CABG). The Ontario Cardiac Care Network (CCN) produces hospital-specific estimates of outcomes including in-hospital and 30-day mortality, and complications, such as hospital length of stay (LOS), ICU LOS and blood transfusion. The models used to create the risk-adjusted hospital-specific estimates are standard logistic or Poisson regression, adjusting for relevant patient risk factors obtained from clinical chart data, with no inclusion of hospital random effects or adjustment for clustering of patients within hospitals. From these models, CCN produces expected numbers of events at each hospital and computes indirectly standardized rates as observed divided by expected rates. The estimates are disseminated to the 14 CABG hospitals and published in a report on the CCN website that requires a member password to view the information (http://www.ccn.on.ca/index.php). See Spencer et al. (2008) for additional details.

G.3 The UK Approach

(Based on comments of an external reviewer)

Methods used by the Care Quality Commission (CQC) are based on using data to target regulatory activity, i.e. giving data to inspectors to guide them in their work such that regulatory activity is targeted. In generating data to do this, a variety of techniques are used, including generating risk estimates that are displayed as a series of dials in the "Quality and Risk Profile" (QRP). See,

```
http://www.cqc.org.uk/guidanceforprofessionals/nhstrusts/
ourmonitoringofcompliance/qualityandriskprofiles.cfm
```

These are used by inspectors and can be used to initiate surveillance techniques. These use time series data and other statistical methods to generate findings where there are higher than expected death rates (known as the outliers program). See,

```
http://intranet.cqc.local/CQCIntranet/news--events/news--updates/
2010/cqc%e2%80%99s-outliers-programme.aspx
```

In addition, the Intelligence Directorate at the CQC employs a range of other techniques and products that use data to guide regulatory activity. These include thematic reviews and a program of National Health Service surveys. In some cases information is used to very quickly trigger activity with a provider (e.g., information received from a whistleblower), while in other cases a statistical model is used to consolidate a large number of data sources using different utilities and weights (the QRP approach).

H Flexible Prior Distributions and Histogram Estimates

Moving away from a single, Gaussian prior distribution for the random effect in a hierarchical logistic regression has the potential to absorb some of the controversy associated with over-shrinkage,

wherein either outliers are masked or high variance (low volume) hospitals are shrunken close to the national value. In addition, using model output to compute SMR estimates with a distribution that better reflects the distribution of the true, underlying SMRs should be considered. For each shrinkage can be less than with the normal without the need to include hospital-level attributes.

H.1 Low degree of freedom t-distribution prior

Use of a low degree of freedom (df), t-distribution prior will avoid over-shrinking truly outlying hospital effects. However, its use will only minimally affect shrinkage for low volume (high variance) hospitals because the degree of shrinkage control depends on the Z-score (estimate/SE) of the MLE estimated hospital effect from a logistic regression. Figure 1 displays the relation among the df, observed data and posterior mean and highlights this Z-score dependence.

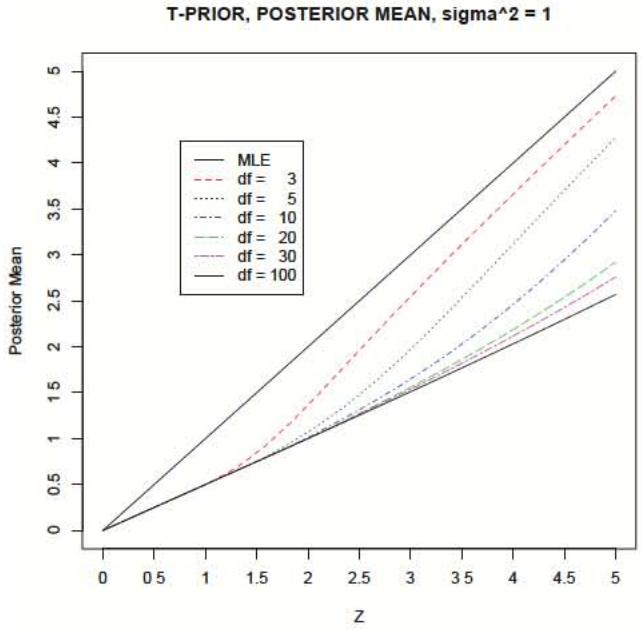

Figure 1: Posterior mean a function of the observed data ($Z = MLE/\sigma$ = MLE/SE) for various degrees of freedom. The priors are scaled to have $\tau^2 = 1$ and for this example the conditional variance $\sigma^2 = 1$. So, for the normal prior (df $= \infty$) the weight on the data is 0.50. The MLE (45° line) is also included.

Note that for 5 df , when the MLE is less than 2·SE from the population value (0 in our example), the posterior mean is virtually identical to that for the normal prior. Beyond that, the t-prior shrinks less and as Z increases the posterior mean is approximately a fixed translation of the MLE. This shrinkage control will maintain identification of outliers (large $|Z|$), but will still impose considerable shrinkage towards the population mean on average for at least 95% of the MLEs (those less than 2 times their sampling standard deviation from the population mean). Thus, much of the stabilization conferred by the normal prior will be retained. Importantly, in addition to

maintaining more weight on the highly deviant MLEs than does the normal prior, the posterior variance associated with a low degree of freedom t-prior will correspondingly increase with the deviation of the MLE from the population mean.

H.2 Mixture of normal or t distributions

As Magder and Zeger (1996) show, a mixture of distributions is an effective way to broaden the family of prior distributions. Relative to a single normal with 2 degrees of freedom (mean and variance), the three-component mixture requires 8 df (2 df for each normal plus 2 df for the mixing weights {the three weights must add to 1.0}), 6 df more than for the single normal. Generally, components of latent mixtures can be poorly identified, but the estimated overall shape will be very stable.

The strategic computations would be identical to the current approach, but the form of the prior would be different (e.g., multi-modal). So, shrinkage would be more complicated than with the single normal, but still there would be a posterior mean. The posterior mean would be a weighted average of the three posterior means associated with the three normal components; the weights would be proportional to the prior odds times the marginal likelihood ratio with marginal likelihoods computed from the normals. With a modest increase in computational overhead, the normal components of the mixture can be replaced by t-distributions.

Though attractive in principle, the mixture approach will have little effect on the estimates for hospitals with a small number of events because for these hospitals the likelihood ratio of the two marginal distributions will be very close to 1.0 and the shrinkage target will be close to the national mean (the *a priori* weighted average of the component-specific means).

H.3 Semi-parametric priors

Paddock et al. (2006); Paddock and Louis (2011) evaluate performance of Dirichlet process priors in estimating the underlying prior (they don't provide information on consequent performance in estimating parameters such as RSMRs). With the large number of hospitals being evaluated by CMS, sufficient data are available to support this approach (for an example, see Lin et al., 2009), but to ease communication, maintain credibility, and likely add sufficient flexibility, use of either a single t-distribution or a mixture of three normal distributions is likely sufficient.

H.4 Use of a prior distribution other than Gaussian

The possibility of replacing the normal distribution provides another entry point to the issue of possible over-shrinkage in the context of assessing hospital quality of care. The prior distribution for the hospital random effects describes the plausible range, mean, mode, and other features of these effects, after adjusting away patient risk factors. Is it sensible to move away from a symmetric, unimodal prior or should it be at least approximately symmetric so that some intercepts are high and some intercepts are low with the median and the mode either identical or very close? In the

context of a symmetric, unimodal prior, does the use of low degree of freedom t-distributions with longer tails provide insufficient stabilization for too many hospitals? What does use of a multi-modal prior imply about computation of risk-standardized mortality rates? These questions relate to the underlying arrangement of true hospital effects and to policy goals. Therefore, discussion and evaluation of these and other options address policy issues, of course as manifested in statistical considerations.

I Histogram estimation

The empirical distribution function (edf) or histogram of estimates based on posterior means of target parameters is under-dispersed relative to that for the true-underlying values and the edf of the direct (MLE) estimates is over-dispersed. Shen and Louis (1998) proposed triple-goal estimates that optimally estimate the edf, produce optimal ranks and lose very little of the estimation advantages conferred by posterior means. Lin et al. (2006, 2009) generalized the approach and applied it to evaluating dialysis center SMRs using the United States Renal Data System database. These estimates are more spread out than the posterior means and should be considered as an alternative to using posterior means. They have the added benefit of compatible results for all monotone transforms of the target parameter. For example, the same estimated SMRs are produced by direct analysis of them or analysis of the log(SMRs) followed by exponentiation.