4. Methods for Assessing the Usefulness of Outcome Reports

To examine the role of random variation versus true hospital quality differences in assessing reported hospital outcomes, we developed simulations to determine how often hospitals would be mislabeled in public reports. To do this, we first made assumptions about what the population of hospitals looks like in terms of both the proportion of hospitals with superior, good or expected, and poor quality and the difference in outcomes between these groups of hospitals. The second step was to calculate, given the first assumptions, the probability that an individual hospital with known characteristics will receive a particular label (e.g., "poor" vs. "good" vs. "superior") and how often those labels will be misapplied (e.g., that a poor quality hospital will be labeled "good"). This mislabeling is possible because random variation in patient outcomes can occur such that, by chance, a good hospital could potentially have a significantly worse than expected mortality rate. How often this happens is a function of the difference in performance rates between good and bad hospitals and the sample size at each hospital (which determines the standard deviation of measured performance for like hospitals).

The starting point for our work was an article by Thomas and Hofer,[8] one of a series from this research group in which they conclude that the inherent random variation in outcomes—that is, the well-recognized phenomenon of variation around an expected mortality rate caused by chance alone and not failures of care or patient risk factors—makes the use of outcome measures for public reporting (and presumably for QBP) misleading and inaccurate. Random variation is important because most outcomes reflect rare events, e.g., a 5% mortality is relatively high for surgical procedures and 15% is high for medical admissions. Also, because most hospitals have relatively small numbers of patients for most conditions and procedures, 200 patients with a given condition is high. Moreover, patients either live or die, so there will be a distribution of mortality rates around the "true" value for a hospital.[66] The question is whether this random variability creates so much "noise" that it is impossible to detect the "signal" indicating truly superior or poor hospitals.

For the sake of simplicity, and because it has been done in much of the prior literature, we focus our analysis below on mortality rates. However, the same concerns about the impact of chance and the same approaches to assessing its impact apply to any of the other major outcomes of interest, from patient satisfaction to complication rates to long-term disability rates and even cost (although the specific statistical approaches are slightly different for continuous variables than for binary variables). With a similar rationale, we focus here on hospitals. Again, the analysis could be applied at other units of observation, such as individual providers, teams, or even health plans.

General Approach to Simulation

In the six scenarios simulated in this report, we refer to each set of underlying assumptions as a *hypothetical world* with known hospital characteristics, recognizing that these assumptions are necessarily simplifications of the real world and are certain to be at least slightly inaccurate. (If, under the given simplifying assumptions the proposed approaches for reporting do not seem to

work, as is argued by Thomas and Hofer, then they are unlikely to work in the more complex real world. On the other hand, if certain reporting approaches seem to work under plausible assumptions, further tests are then warranted to make sure they are still valuable under more realistic situations.)

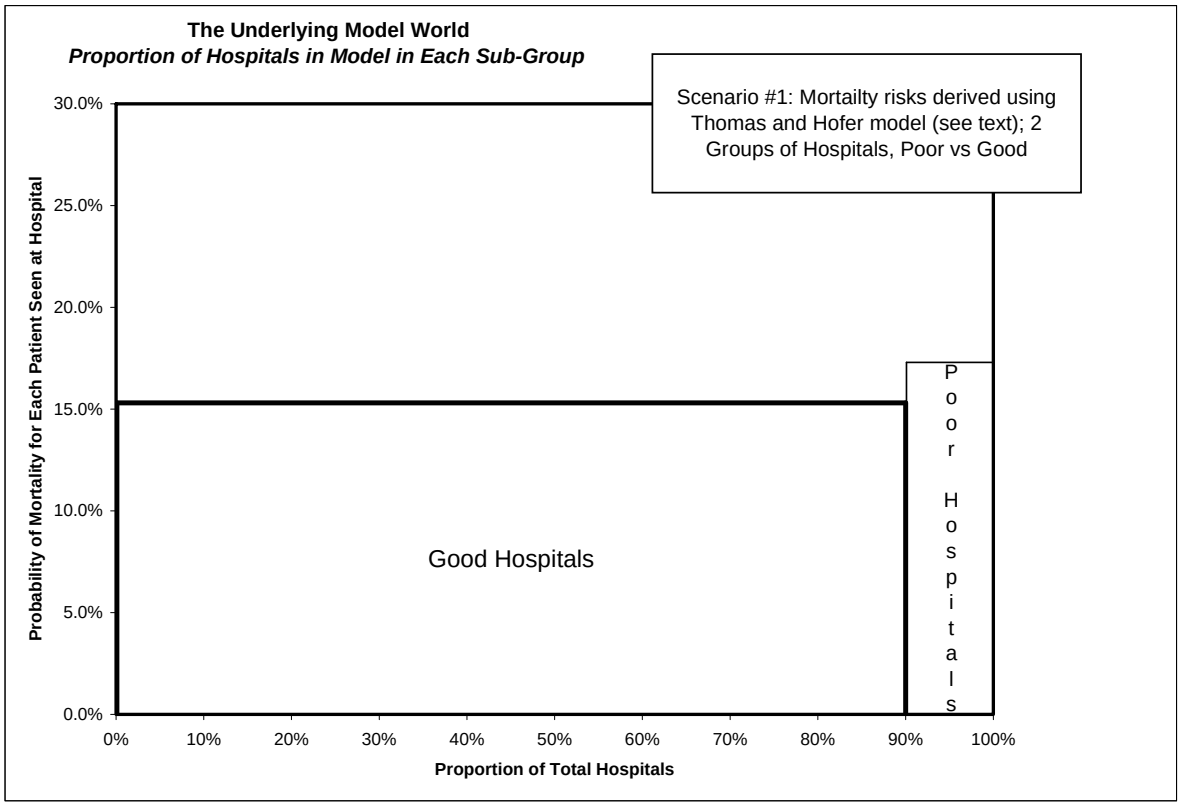
Given the assumptions made in each scenario, we then apply a performance label to each individual hospital (e.g., "poor" vs. "good" vs. "superior"). We refer to this labeling process as the *evaluation system*, and the frequency of mislabeling is determined both by the assumptions about the hypothetical world and by the approach to evaluating hospitals. The design of an evaluation system is not a purely statistical question—it also reflects how the labels are to be used. Thus, if the label is intended to be used by itself in front page headlines one may reasonably want to be much more sure of its accuracy than if it is seen as one of many indicators that needs to be confirmed with detailed chart reviews.

The hypothetical model is a simplified representation of what the world of hospital quality actually looks like. By varying our assumptions over a reasonable range of values, we can determine the robustness of the evaluation system. In the application of evaluations to real-world hospital outcome data, one would not know which hospitals were actually poor or good in advance. One would only be able to observe the measured performance, such as mortality rate, from each hospital. It would be the job of the evaluation system to assign each hospital a label, which would hopefully reflect the true nature of the hospital's performance. However, each hospital's outcomes in any given year are affected by chance; a patient may receive perfect care and die anyway; another patient may receive poor quality care yet survive. On average, though, we would expect higher death rates in poor quality hospitals.

In Thomas and Hofer's hypothetical world (scenario 1 below) there are only two types of hospitals. Poor quality hospitals comprise 10% of all hospitals, and good quality hospitals account for the remaining 90%. The defining difference between them is the proportion of patients receiving "good processes of care" and "poor processes of care" at each hospital in each group. Thomas and Hofer apply data from the literature and a program of chart reviews of implicit quality of care in Texas in 1990 and 1991 to make a series of calculations to determine the average risk of death per patient receiving care at each type of hospital. The input parameters which feed into their model of the hospital world include the risk of death having received good care, the risk of death having received poor care, the odds of receiving poor care at a good hospital versus a poor hospital, the number of patients at the average hospital, and the proportion of hospitals that are *poor*, as defined above. In their model, the difference in overall mortality rates between *good* and *poor* hospitals is very small (15.3% vs. 17.3%), so it is not surprising that they find it difficult to label hospitals accurately due to the effects of random variation.

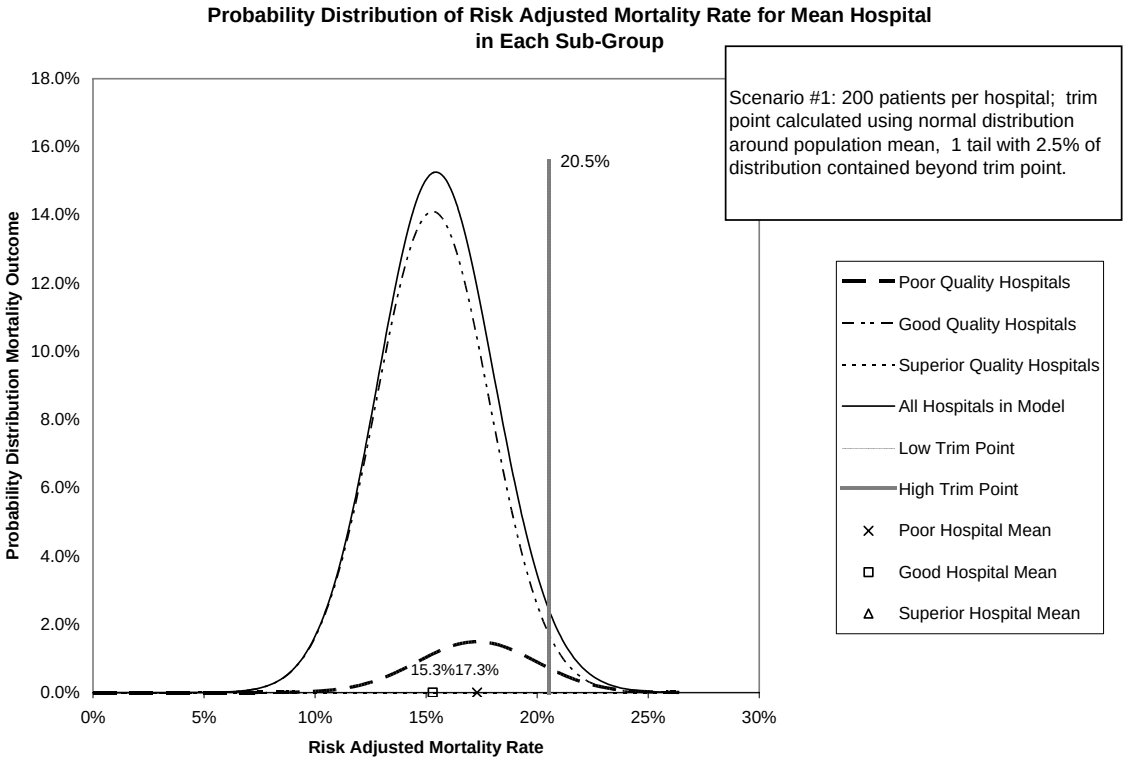
A graphical representation of this hypothetical world of hospitals is shown in Figure 5.

Figure 5: Hypothetical world of hospitals



To label hospitals, Thomas and Hofer used an evaluation system similar to clinical diagnostic tests. They defined poor performance as that which would be found in the high mortality tails of a distribution normally distributed about the mean hospital performance. In their trials, they used a 5% cutoff, so performance likely to occur by chance in only 5% of situations was labeled as being an "outlier." As outliers can occur both in the poor performance tail, and in the superior performance tail, only 2.5% of hospitals would be labeled "poor." The value for mortality data, above which 2.5% of hospital performance would be expected to fall is called the high trim point.[8] The evaluation system is summarized graphically in Figure 6, which is adapted from Thomas and Hofer.

Figure 6: Hypothetical world and evaluation function (adapted from Thomas and Hofer[8])



In summary, the evaluation system inputs are only the mean performance of hospitals (something observable), the number of patients seen in each hospital, and a given year's mortality data for the particular hospital. With these data, the evaluation system generates a label of "poor quality" if the mortality rate of the given hospital is greater than the trim point and "good quality" if the result is less than the trim point. Note that this approach simulates the real world in which an evaluator tries to grade hospital outcomes given only the hospital performance data. He/she does not know *a priori* which hospitals truly have poor or good quality. That is, only the summary solid curve describing the observed mortality rates for *all* hospitals in Figure 6 and the trim point are known; the dashed lines are not known in the real world, but are used only to create the hypothetical world, upon which the grading function is tested. Furthermore, there may not be data from the hundreds or thousands of hospitals needed to plot the type of smooth solid curve shown. Instead, one may merely have a good estimate of the overall risk-adjusted mortality rate and then assume a normal distribution.

Enhancements to the Thomas and Hofer Model

In our simulations, we enhanced the Thomas and Hofer approach in three ways. First, we increase the sophistication of the assumptions about what the underlying hospital population looks like, allowing for the existence of hospitals with superior quality and drawing our estimates of the percentage of "poor", "good", and "superior" hospitals from more recent data. We then consider alternative assumptions for input parameters for the evaluation system and use

more sophisticated grading functions—including multi-category grading and evaluation over time.

The first enhancement to the Thomas and Hofer model investigated was the addition of a third sub-group: "superior quality hospitals." Based on published California data from 1996-1998 showing approximately 10% of hospitals had been labeled "worse than expected" and 10% had been labeled "better than expected", we altered the hypothetical world of hospital performance to include 10% poor quality, 10% superior quality, and 80% good or expected quality hospitals. Furthermore, hospitals labeled "better than expected" had been shown in validation studies to have superior processes of care compared to hospitals labeled "worse than expected". Thus, although a simplification (hospital performance is likely aligned along a spectrum, rather than divided into only three groups), these results support the assumption of a distribution of hospital performance that included 10% poor quality, 10% superior quality, and 80% good (or expected) quality hospitals.[67, 68]

We obtained estimates of probability of death at poor, good, and superior quality hospitals using three-year grouped data published in the California study of acute myocardial infarction outcomes.[67, 68] Hospitals that were consistently—over two or three studies—i.e. six or nine years—found to be statistically significantly better than the mean performance of California hospitals were included in the group of superior hospitals. Those hospitals with consistent performance below the mean were used to form the poor group. The remaining hospitals—those whose performance was not consistently and statistically different from average over two or three study periods—formed the "good" or "expected" group. The characteristics of these groups are shown in Table 13, Scenarios 3 through 6.

We believe these assumptions are a reasonable starting point for building a hypothetical world of truly poor, good, and superior hospital quality. We assume that the risk adjustment model used in the California report does not have substantial biases. Additionally, hospitals labeled "better than expected" were found in validation studies to have superior processes of care compared to hospitals labeled "worse than expected."[69]

Changes were then made in the evaluation or scoring system used to label a set of outcome results as either "superior," "good," or "poor." We assessed the accuracy of labeling using two tailed outliers, so that we could recognize and label hospitals with superior outcomes (i.e. hospitals with measured risk adjusted mortality below the trim point are labeled "superior") as well as those with poor outcomes. We then repeated these assessments with different outlier trim points—trimming from 2.5% - 10% into each tail, such that with two tailed trim points, either 5% or 20% of hospitals would be labeled as either "poor" or "superior." We also ran simulations using 1, 2, and 3-year evaluations, such that each hospital would receive labels for each of 3 years. The sum of the annual grades over the 3-year period would serve as a "meta-score." For simplicity, a *star* system was employed, in which a grade of "poor" was assigned *1 star*, a grade of "good" received *2 stars*, and a grade of "superior" earned *3 stars*. The minimum 3-year score for a given hospital is therefore *3 stars* (obtained by receiving only 1 star in each of the 3 years); the maximum is *9 stars*.

To calculate multiple year probabilities, the probability for each score for one year was calculated for each hospital group as described above. Then, all possible combinations (order not important) of grades for 2 or 3 years was enumerated, and the cumulative probability that a given number of each grade was assigned was calculated by multiplying the appropriate probabilities for each grade. The results were then tabulated by hospital group (corresponding to sensitivity and specificity measures) and then by score assigned (corresponding to predictive errors).

Table 13 summarizes the six scenarios to be simulated. (See Appendix B, available at www.ahrq.gov/clinic/epcindex.htm, for the simulation algorithm.)

Table 13: The six scenarios simulated

Scenario #	Hypothetical (Defined) World of Hospitals							Grading Function		
	Superior Quality		Good Quality		Poor Quality		Average Number of Patients per Hospital	Mean probability mortality of whole population	Low Trim Point < Labeled superior	High Trim Point > Labeled poor
	True Probability of Mortality	% Total Hospitals	*True* Probability of Mortality	% Total Hospitals	*True* Probability of Mortality	% Total Hospitals				
1	Only 2 Groups		15.3%	90%	17.3%	10%	200	1 tail distribution: grade is either "good" or "poor", i.e. if outcome is > high trim point, which includes 2.5% of population		
	Recreation of Thomas and Hofer model, as starting point.							15.5%	N/A	20.5%
2	13.3%	10%	15.3%	80%	17.3%	10%	200	2 tails: with ~2.5% of population above/below each;		
	Thomas and Hofer model; now with three groups; mortality rate for "superior" calculated using assumption that superior hospitals are as much better than good quality hospitals as poor quality hospitals are worse than good quality hospitals (i.e. rate at superior hospitals = rate at good quality hospitals – (rate at poor quality hospitals – rate at good quality hospitals); also assume 10% of hospitals are superior quality.							15.3%	10.3%	20.3%
3	8.6%	10%	12.2%	80%	17.1%	10%	200	2 tails: with ~2.5% of population above/below each; mortality outcomes above high trim point labeled "poor," below low trim point labeled "superior."		
								12.1%	7.6%	16.6%
	Mortality values from California AMI study (see text), using Thomas and Hofer hospital group proportions.									
4	8.6%	10%	12.2%	80%	17.1%	10%	100	2 tails: with ~2.5% of population above/below each		
								12.1%	5.7%	18.5%
	As above except number of patients per hospital = 100									
5	8.6%	10%	12.2%	80%	17.1%	10%	100	2 tails: with ~10% of population above/below each		
								12.1%	7.9	16.3
	As above; number of patients per hospital = 100									
6	8.6%	10%	12.2	80%	17.1	10%	400	2 tails: with ~10% of population above/below each trim point.		
								12.1%	10.0%	14.2%
	As above; number of patients per hospital = 400									