Appendix A. Quality-based Purchasing Technical Expert Panel and Peer Reviewers

David Atkins
Chief Medical Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Anne-Marie Audet
Assistant Vice President, Quality Improvement
The Commonwealth Fund

John Bott *
Value Based Purchasing Manager
Employer Health Care Alliance Cooperative

Douglas A. Conrad
Director, Center for Health Management Research
University of Washington

Janet Corrigan
Division of Health Care Services
Institute of Medicine

Judith Hibbard
Professor, Department of Planning, Public Policy and Management
University of Oregon

Donna Marshall *
Executive Director
Colorado Business Group on Health

Peggy McNamara
Senior Analyst
Center for Delivery, Organization and Markets
Agency for Healthcare Research and Quality

Arnold Milstein *
Managing Director
William M. Mercer

Ann Robinow *
President and Chief Executive Officer
Patient Choice Healthcare

Dennis Scanlon
Assistant Professor
Department of Health Policy and Administration
The Pennsylvania State University

Stephen Schoenbaum *
Senior Vice President
The Commonwealth Fund

Laura Tollen *
Senior Policy Consultant
Kaiser Permanente Institute for Health Policy

*Member of Technical Expert Panel

Appendix B: General Approach to Simulations

The algorithm for each simulated scenario is as follows:

1. Create a hypothetical hospital world based on input parameters using data available from the real world. These models contain either two or three homogenous groups of hospitals each with a defined level of hospital performance. This model is the world of true hospitals, or the gold standard of the model.

Our hypothetical model is somewhat conceptually different from Thomas and Hofer's. Instead of using the likelihood of receiving poor or good quality care, we differentiated hospitals based on the overall level of care provided to *all* patients. A good hospital may have processes or personnel in place to provide better quality care to each of its patients, not just to limit poor care to fewer of its patients. This assumption allows us to build a world view identical to Thomas and Hofer, but to start deeper in their model, at the level of probability of death in each hypothetical hospital group (without deriving these values from their assumptions outlined above).

2. Apply a grading function to a set of hospital outcomes. In our simulation outlier cutoffs, or "trim points," were used to label outcomes as "poor," or "good," or in models with three categories, "superior." The value of the trim point is estimated by assuming that the observed mortality risk outcomes assume a normal distribution around the mean mortality rate of the hospitals. The trim point(s) are set such that a given percent of the mortality outcomes of the population of hospitals will fall above or below the respective poor and superior trim points. Other possible grading functions could use arbitrary trim points (for absolute standards of quality), trim points based on reference populations, or trim points based on other distributional assumptions.

Note that the Thomas and Hofer evaluation function assumes that the overall distribution – that which can be observed, is equivalent to a normal distribution around the mean hospital probability of death, with standard deviation defined using the number of patients at each hospital. In reality, the sum of the "good" and "poor" distributions – the solid line in figure 2, is actually a right skewed distribution, due to the larger standard deviation of the "poor" sub-group, as a function of the higher probability of mortality in this subgroup, as calculated with the following equation: std_dev of poor group = Squareroot (prob_death * (1 – prob_death)/num_patients_per_hospital). Note also that these distributions are not truly normal, as they terminate at 0.0 (i.e. there is no negative probability of death).

3. Assess the performance of the evaluation system – either via sensitivity and specificity (i.e. how likely is the system to correctly label poor quality hospitals as "poor" and superior quality hospitals as "superior") or predictive values (i.e. given a grade of "superior," how likely is a hospital actually to be of superior quality?). The former measure is of most concern to hospitals, concerned about being mislabeled, while the accuracy of predictive values tells consumers, purchasers, and other policymakers how much to trust the grades assigned. The perfect evaluation system would label each hospital according to the true world group to which it belongs.

This step is repeated for a given grading function over several possible hypothetical hospital worlds (see step 1) to test the robustness of the evaluation system. Results from the representative scenarios are discussed in Section 3.

The models were produced using Microsoft Excel with statistical functions and Visual Basic for Applications, 2003. Each parameter was either entered by hand, or derived using a recreation of the Thomas and Hofer model or from empiric data as described above. For each hospital group, the chance of each grade was determined using the NORMDIST function, which given a mean (in this case, the mortality risk as defined for the group), standard deviation (calculated using the group's mortality probability and number of patients per hospital), and a trim point (the trim point as defined in the approach to evaluation and labeling, based on the observed, total distribution of hospital mean mortality), returns the probability of selecting an outcome that exceeds the trim point, assuming a normal distribution based on the mean and standard deviation supplied. This corresponds to the area under the hospital group's curve that is to extreme side of the trim point line.

Appendix C:

Assessing the Usefulness of Outcome Reports

In this appendix, we review the methods and results of all the simulations performed in full detail. Some of the figures and tables are the same as those already presented in the body of the report.

Methods for Simulations

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. We sought to assess how the frequency of mislabeling depended upon (a) underlying assumptions about the true differences in hospital quality and (b) different evaluation and labeling strategies. The starting point for our work was an article by Thomas and Hofer,[1] 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.[2] 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).

General Approach to Simulation

In simulating the use of outcomes data for QBP, there are two distinct steps to assessing the impact of random variation on reported hospital performance. The first is to choose assumptions about what the population of hospitals looks like in terms of both the proportion of hospitals with good and poor quality and the difference in outcomes between these groups of hospitals. In doing this, we are assuming 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.)

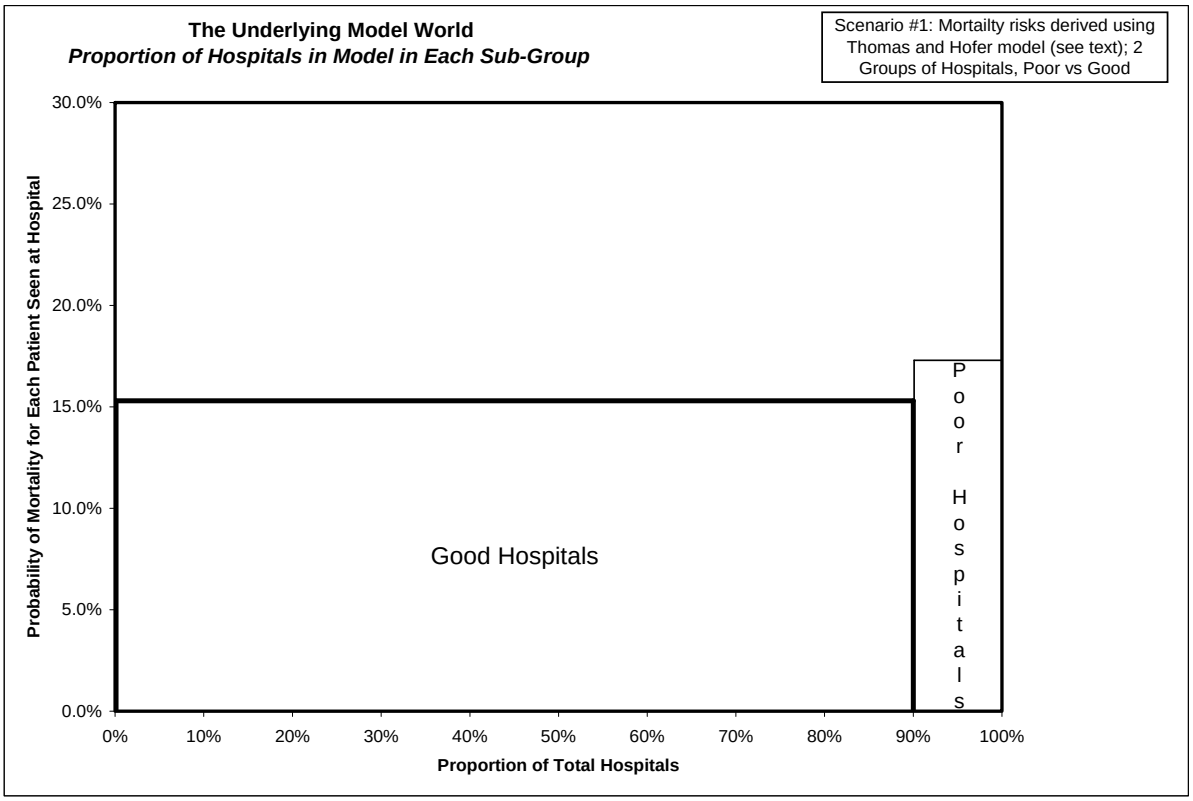
The second step is to calculate, given the first assumptions, the probability that an individual hospital with known characteristics will receive a particular label (e.g., "poor" vs. "good" vs. "superior") and how often those labels will be misapplied (e.g., that a poor quality hospital will be labeled "good"). We refer to this second function 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 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—qualitatively—poor or good in advance. The input to the evaluation system would only be 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. 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, however, we would expect higher death rates in poor quality hospitals.

In Thomas and Hofer's model, the hypothetical world of hospitals is composed of two groups.[1] 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 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 C 1.

Figure C 1: Hypothetical World of Hospitals

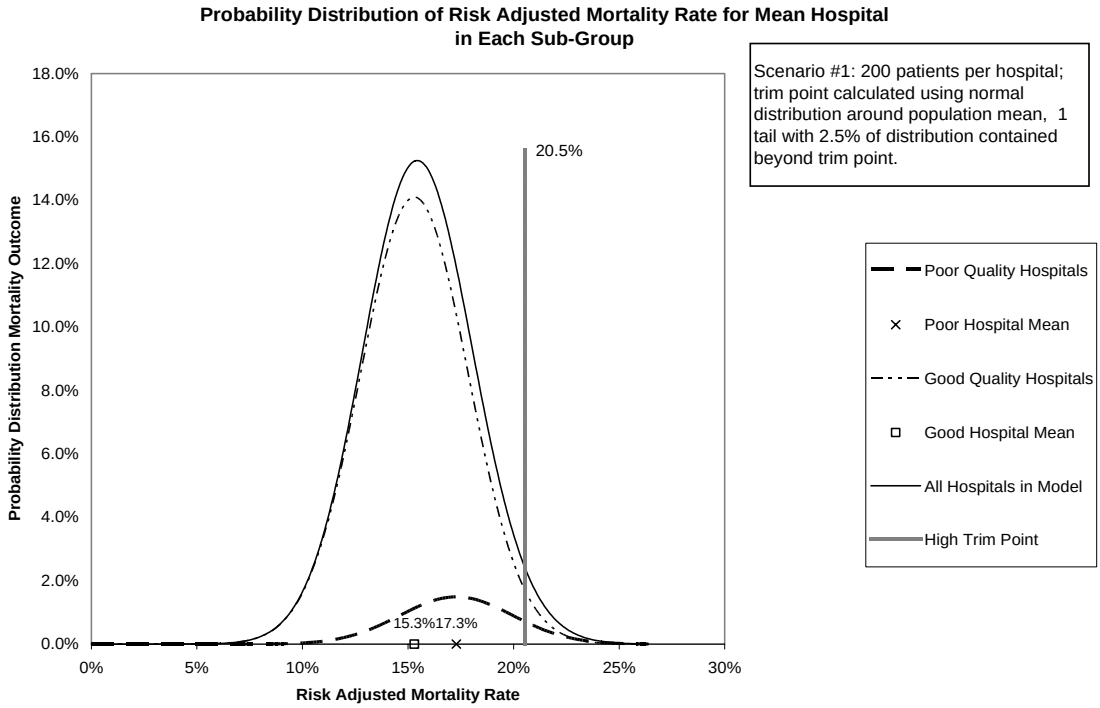


To label hospitals, Thomas and Hofer use an evaluation system similar to clinical diagnostic tests. They define poor performance as that which would be found in the tails of a distribution normally distributed about the mean hospital performance. In their trials, they used a 5% cutoff, so performance 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 occur is called the high trim point.[1] The evaluation system is summarized graphically in Figure C 2, which is adapted from Thomas and Hofer.

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 C 2 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.

Figure C 2: Hypothetical World and Evaluation Function (adapted from Thomas and Hofer[1])



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.[3,4]

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.[3,4] 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 C 1, 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."[5]

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 C 1 summarizes the six scenarios that will be simulated.

Table C 1: 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									

Results of Simulations

Scenario 1: Reproducing Thomas and Hofer

For this scenario, we reproduced in our model the assumptions of Thomas and Hofer. The probability of death at *poor* and *good* hospitals was calculated as in their model as described in an unpublished appendix to their paper. The scenario is summarized by Figure C 1 and Figure C 2 above, and Table C 2 and Table C 3, below.

Notice that in this scenario, a fairly large part of the *poor* quality hospital distribution is intersected by the trim point (Figure C 2). Examining the areas under the *good* quality and *poor* quality hospital curves, to the right of the trim point, it appears that some hospitals that are labeled *poor*, may in fact be of *good* quality. This error is called predictive error, and is reported in Table C 2. Other predictive values—positive predictive value (the chance that a hospital which received a *poor* grade is actually a *poor* quality hospital) and negative predictive value (the chance that a hospital receiving a *good* grade is actually a *good* quality hospital)—are shown as well. In the calculation of predictive values, the proportion of the two populations is important. The more rare the condition or state of being "positive" is (in this case, being a *poor* quality hospital), the higher the positive predictive value will tend to be. Since the *poor* quality hospitals only comprise 10% of the population, and their distribution is nearly subsumed by the *good* quality hospitals, it is not surprising that the positive predictive value is so low, and the inversely-related predictive error is so high.

Table C 2: Scenario 1: Predictive Values, Year 1

Score assigned	Hospital *really* is	Probability in whole distribution	Probability within this group of scores	2 category test clinical test labels
Poor	Poor	1.1%	**38.7%**	**Positive Predictive Value**
	Good	1.8%	**61.3%**	**Predictive Error**
	Subtotal	***2.9%***		
Good	Poor	8.9%	**9.1%**	
	Good	88.2%	**90.9%**	**Negative Predictive Value**
	Subtotal	***97.1%***		

Other metrics of test performance are sensitivity and specificity. The measures are independent of the population (or, in this case, hypothetical world of hospitals) in which they are used. They are measures of the tests themselves, and can be used to compare one test with another. To calculate sensitivity and specificity, a gold standard measure must

be used to identify a priori the group to which the individual or organization tested in fact belongs (in our case, as the hypothetical world is defined by us, the gold standard measure is simply the hypothetical world groupings). Table C 3 shows sensitivity and specificity for scenario 1.

Table C 3: Scenario 1, Year 1: Sensitivity and Specificity Calculations

Hospital *really* is...	Score assigned	Probability in whole distribution	Probability within this group of hospitals	2 category test clinical test labels
Poor	Poor	1.1%	**11.2%**	**Sensitivity**
	Good	8.9%	**88.8%**	
	Subtotal	***10.0%***		
Good	Poor	1.8%	**2.0%**	
	Good	88.2%	**98.0%**	**Specificity**
	Subtotal	***90.0%***		

We can see that while the evaluation function will correctly label 98% of *good* hospitals as *good*, it will detect only 11.2% of *poor* quality hospitals in any given year, using Thomas and Hofer's assumptions.

Assessing the Evaluation System over Multiple Years of Use. The results for calculating *star* scores for 2 years are shown in Table C 4 and Table C 5. While predictive values, sensitivity, and specificity are generally defined for tests/functions with dichotomous results, the approach of each can be used with more than one possible outcome. We will examine the predictive value and sensitivity and specificity of the most extreme grades: *2 stars* and *4 stars* over 2 years.

Table C 4: Scenario 1: Probability, Given that a Hospital Has Received Two, Three, or Four Stars over 2 Years, that It is Good vs. Poor

Number of stars (over 2 years)	Probability of actually being poor is...	Probability of actually being good is...	Overall probability of receiving score
2	78.2%	21.8%	**0.2%**
3	36.4%	63.6%	**5.4%**
4	8.4%	91.6%	**94.4%**

For example, the positive predictive value of *2 stars* is 78.2%—a large improvement over the 1-year figure of 38.7%, although only a small set of hospitals will be assigned this grade (0.2%); *4 stars* has a negative predictive value of 91.6%; *3 stars* has poor discrimination between subgroups, although a hospital in this group is more than three times more likely to truly be poor than if one selected a hospital without any performance information (this would be essentially random and would have a 10% chance of yielding a poor hospital, since they are 10% of the general population, but 36.4% of the population receiving 3 stars).

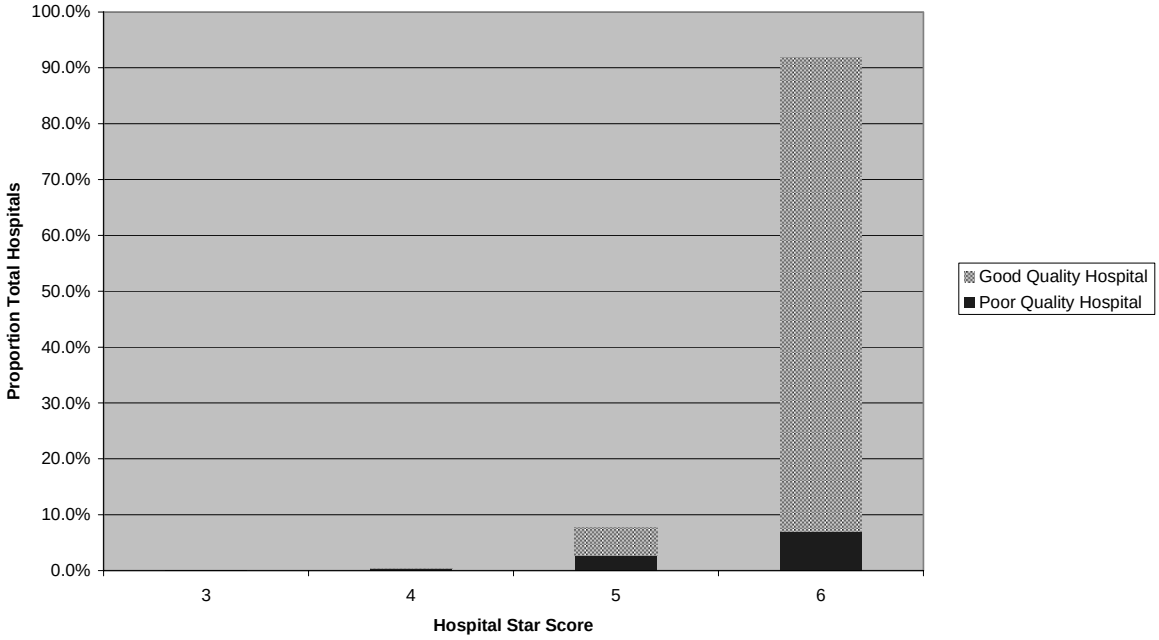
Sensitivity and specificity calculations show that specificity of *4 stars* is 96.1% and sensitivity of *2 stars* is only 1.2%, as 2 stars is very unlikely in this scenario, whether the hospital is poor or good.

Table C 5: Scenario 1: Expected Score Distribution over 2 Years

	Probability (%) hospital will receive score of...			
Hospital *really* is…	2 stars	3 stars	4 stars	Overall probability of being in this group
Poor	1.2%	19.8%	78.9%	**10.0%**
Good	0.0%	3.8%	96.1%	**90.0%**

The results for 3 years of testing in this scenario are shown graphically in Figure C 3 and by hospital group in Table C 6.

Figure C 3: Scenario 1: Percentage of Good vs. Bad Hospitals by 3-Year Star Score

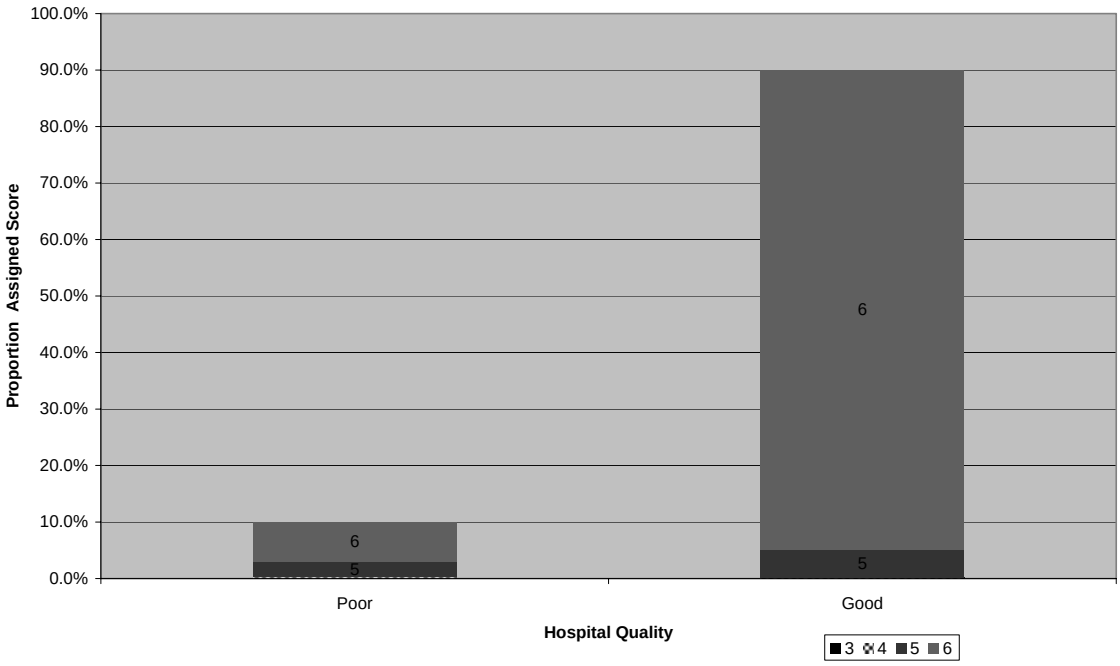


Hospitals with *3* or *4 stars* are almost certainly of *poor* quality—but these scores are rare. Indeed, it is a rare thing to be graded *poor* in this scenario, and to have it occur even once in 3 years happens for only 8.2% of hospitals.

Table C 6: Scenario 1: Expected Score Distribution for Good vs. Poor Hospitals over 3 Years

Hospital *really* is…	Probability (%) the hospital will receive score of...			
	3 stars	4 stars	5 stars	6 stars
Poor	0.1%	3.3%	26.4%	70.1%
Good	0.0%	0.1%	5.7%	94.2%

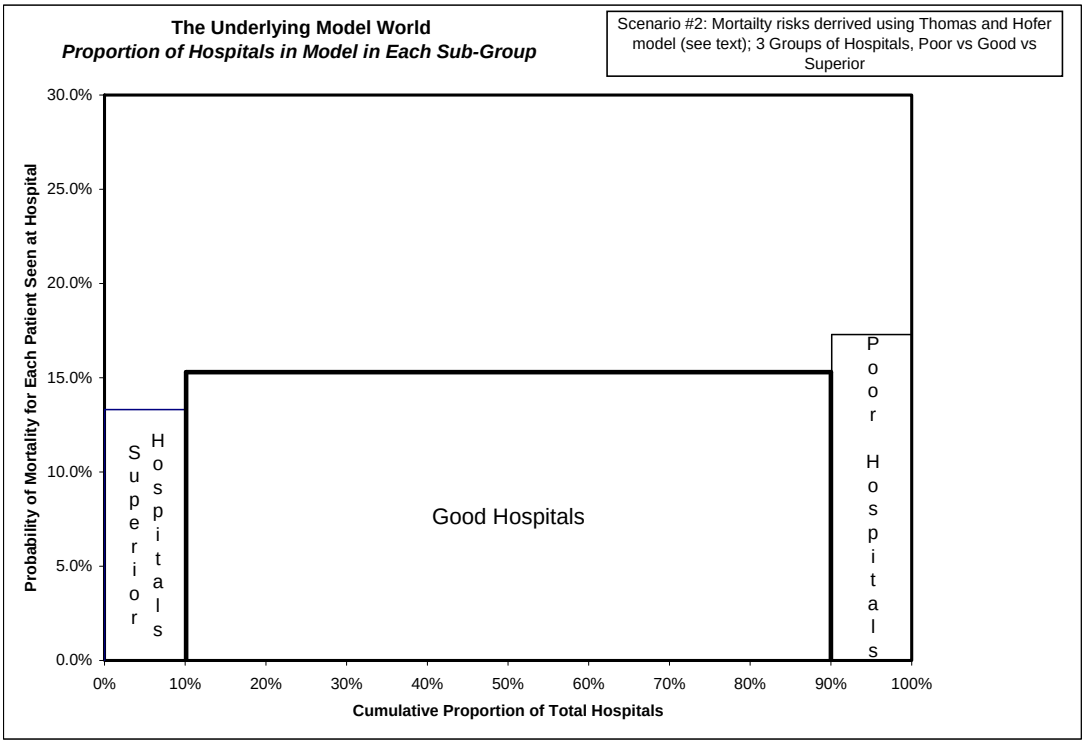
Figure C 4: Scenario 1: Expected 3-Year Score Distribution for Good vs. Poor Hospitals



Scenario 2: Adding Another Hospital Category

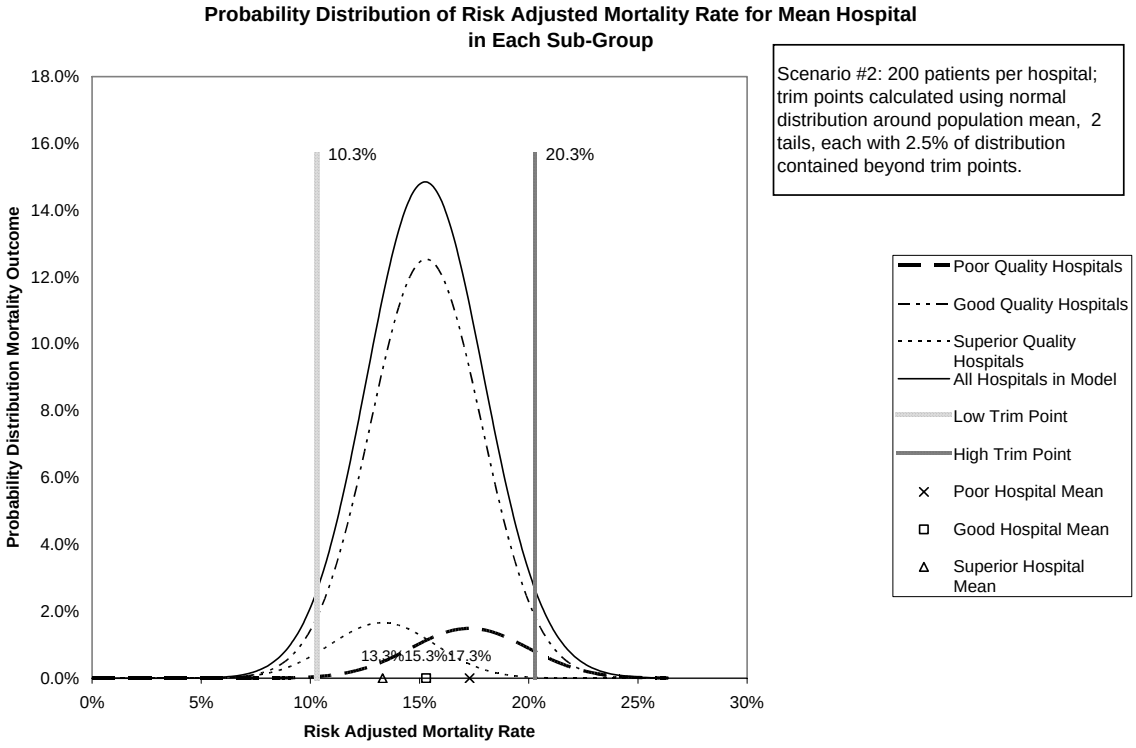
For this scenario, we added the *superior* quality hospital group as 10% of the hypothetical hospital population. The average mortality rate for *superior* hospitals was assumed to be the same percentage difference below the mean performance as Thomas and Hofer's *poor* quality hospitals were above the mean (Table C 1). The mortality rates are shown in Figure C 5.

Figure C 5: Scenario 2: Hypothetical World



The trim points were calculated using the normal distribution based on the average mortality rate with trim points defined so that 2.5% of hospitals would lie under the curve beyond each trim point (in a normal distribution with standard deviation defined by the number of patients per average hospital: 200). These assumptions about trim points and populations are shown graphically in Figure C 6.

Figure C 6: Scenario 2: Hypothetical World and Evaluation Function



Year 1 results now do not have two-value predictive values, sensitivity, and specificity. Instead, the analogous computations are made by score (for predictive values) or by hospital sub-group (for sensitivity and specificity probabilities).

In the 2-year analysis (see Figure C 7), we see that hospitals earning *5 or 6 stars* are all *good* or *superior* quality hospitals. The score of *4 stars* is likely to include hospitals of all types. Low scores eliminate the possibility that the graded hospital is *superior*. However, since nearly 90% of hospitals receive *4 stars*, this evaluation system does not discriminate well among the majority of hospitals.

Three-year *star* scores (see Figure C 8) again reliably identify a handful of hospitals at the extremes of mortality scores. The score of *6 stars* occurs 82.6% of the time, and still includes most of the *poor* and *superior* quality hospitals, as well as a large majority of the *good* hospitals. So, while repeating the scores allows for excellent discrimination of a small number of hospitals (that is, those few with extreme scores have a high chance of being *poor* or *superior*), the large majority of hospitals are still not reliably distinguished from average performance.

Figure C 7: Scenario 2: Proportion of Superior, Good, and Poor Hospitals by 2-Year Star Score

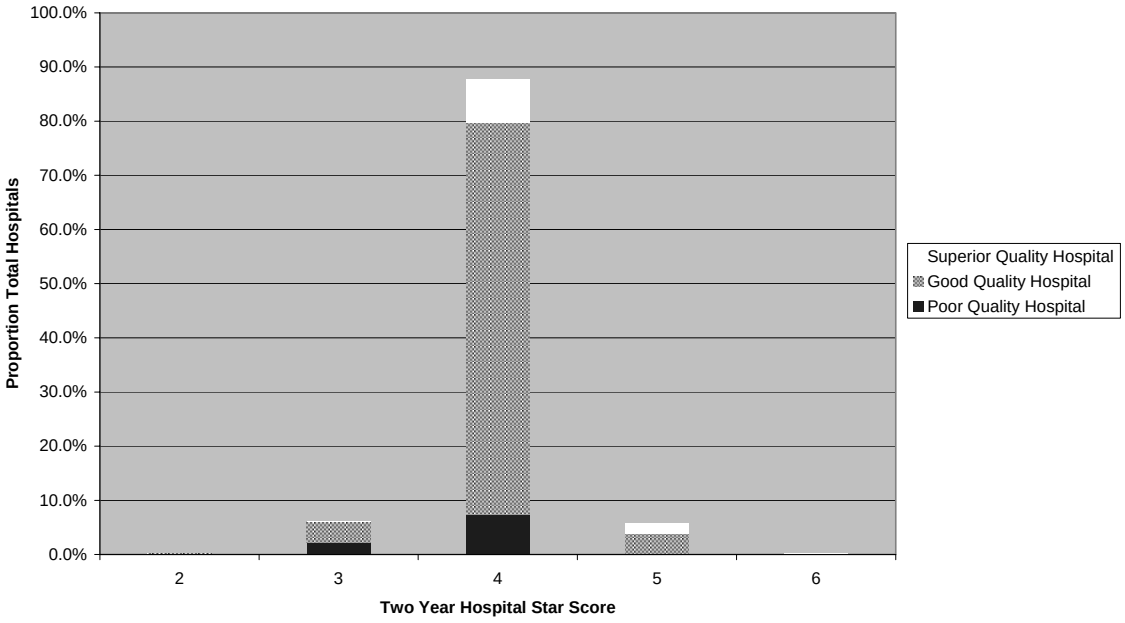
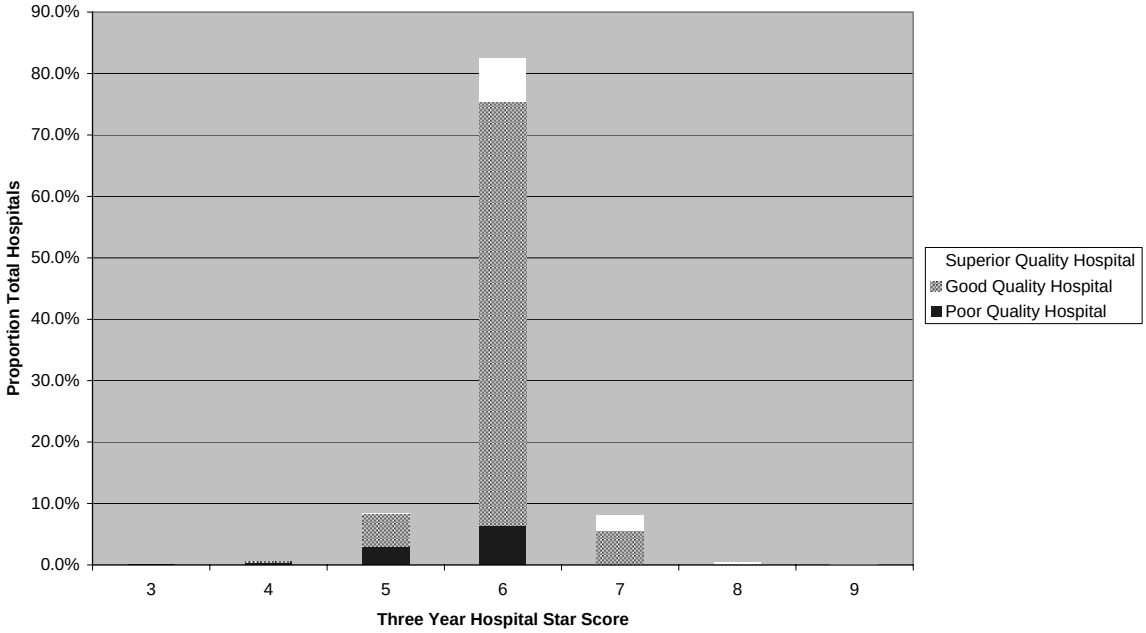
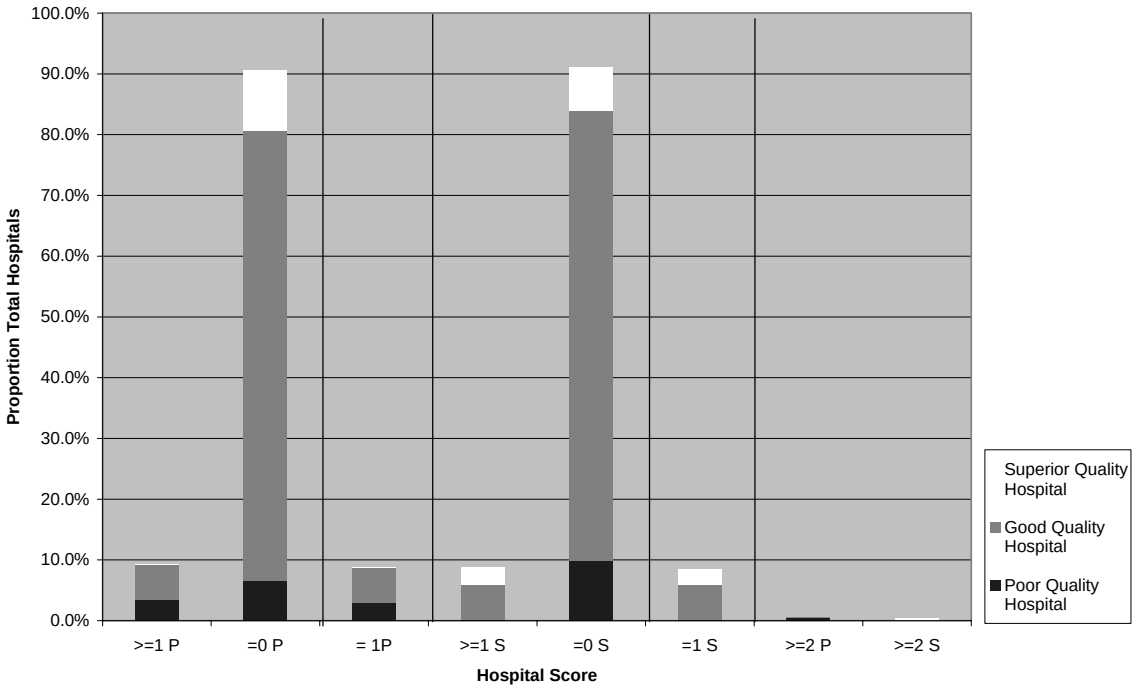


Figure C 8: Scenario 2: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score



Derivative scores were used to assess whether further discrimination could be obtained among the three sub-groups. The measures are *never poor* (= 0 P), *ever poor* (>= 1 P), *exactly 1 poor* (= 1 P), *mostly poor* (>= 2 P), *never superior* (= 0 S), *ever superior* (>= 1 S), *exactly 1 superior* (= 1 S), and *mostly superior* (>= 2S). The derivative scores for scenario 2 are shown in Figure C 9**.**

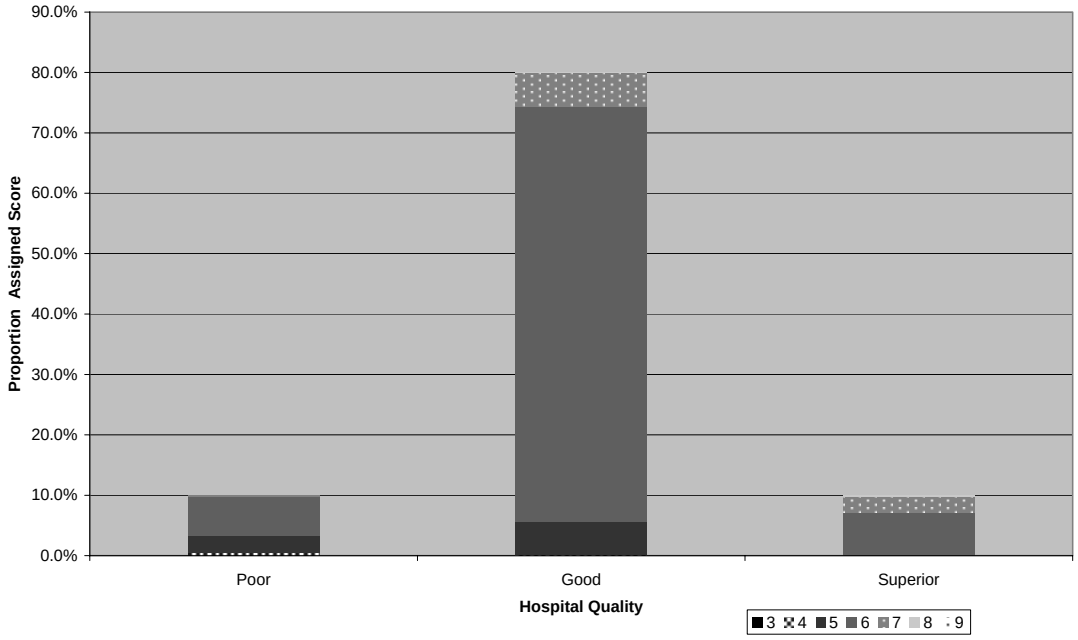
Figure C 9: Scenario 2: Proportion of Poor, Good, and Superior Hospitals with Each Type of Derivative Score



The *ever poor* and *ever superior* scores do eliminate the superior and poor quality hospitals, respectively. However, these scores do not discriminate well between poor and good, or superior and good, respectively. *Mostly poor* and *mostly superior* have high discrimination, but only a trivial number of hospitals actually receive these grades.

Scores for each given hospital group are also summarized in Figure C 10. These results are analogous to sensitivity and specificity calculations for two value evaluations. These results show that *poor* hospitals generally receive scores below *7 stars* and superior hospitals receive *6 stars* or greater.

Figure C 10: Scenario 2: Expected Distribution of 3-Year Star Scores by Hospital Type



Analysis of scenario 2 demonstrated that there could be some improvements to the labels generated by the evaluation system through the addition of multiple year scoring, and more subgroups, and therefore grading categories. However, the underlying hypothetical world has such great overlap between the two relatively rare outcomes of *superior* or *poor* quality, that discrimination is almost by definition difficult. The next scenarios explore using more realistic assumptions about variation in hospital performance to generate the hypothetical world.

Scenario 3: Updating Assumptions about the Hypothetical Distribution of Hospital Quality

For this scenario, the underlying hypothetical hospital model used mortality data obtained from the 1996-1998 California study of risk-adjusted mortality from acute myocardial infarction.[3,4] See Appendix B for the algorithm used to generate the mean mortality for each group.

The model world is shown in Figure C 11 and the evaluation function is summarized in Figure C 12. The evaluation function is based on the reported population mean mortality rate and 2.5% trim points, as described above.

Figure C 11: Scenario 3: Hypothetical World

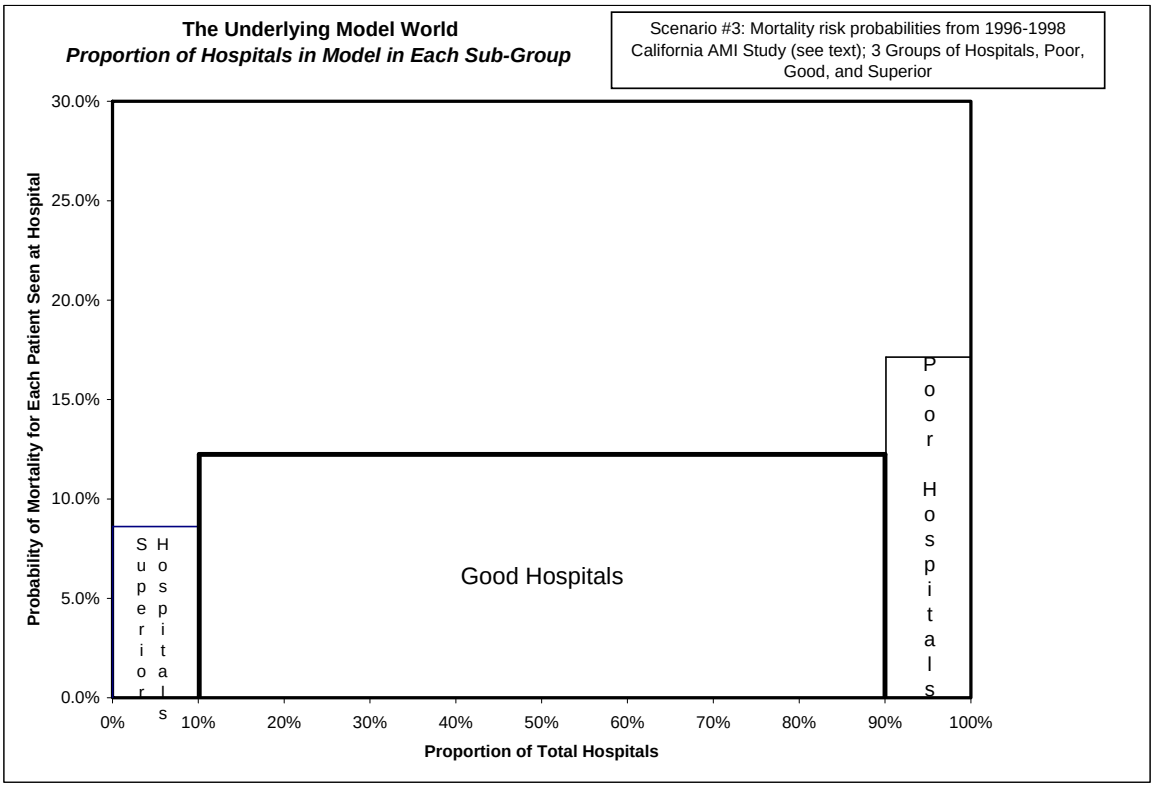
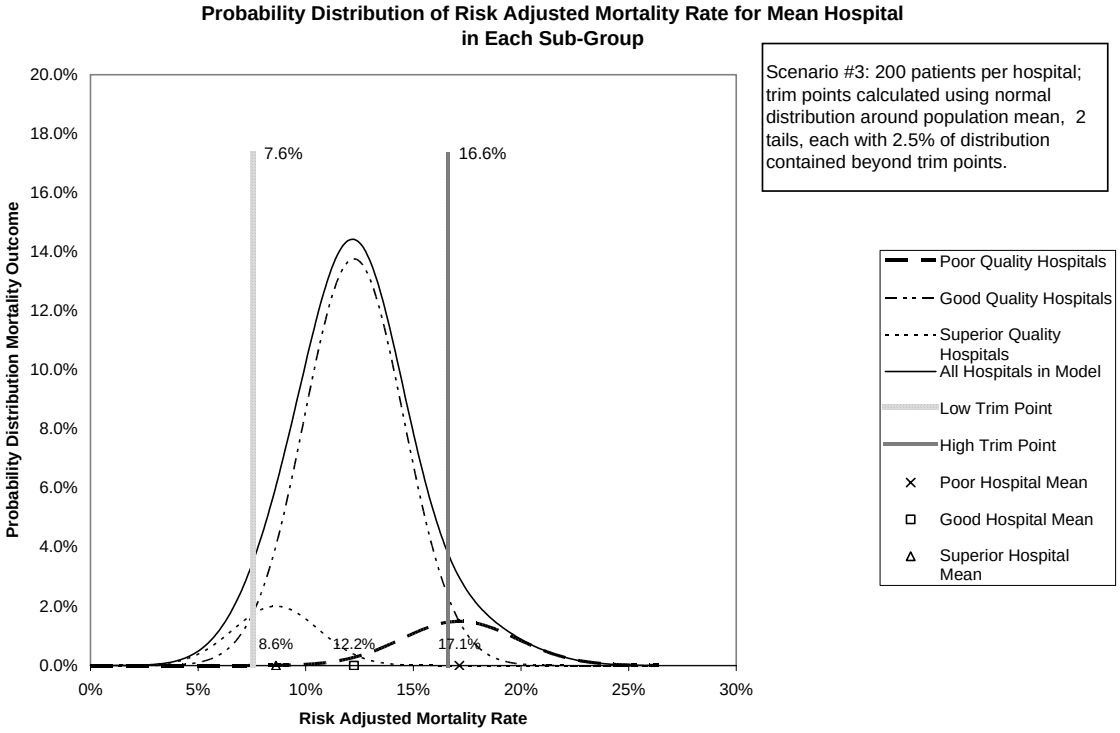


Figure C 12: Scenario 3: Hypothetical World and Evaluation Function



The greater difference between mortality rates in the *superior* and *poor* groups has resulted in better discrimination in 2 year scores (see Figure C 13). A large majority of *poor* hospitals have scores of *2 or 3 stars*, while many *superior* hospitals receive scores of *5 or 6 stars*, and these extreme scores effectively eliminate hospitals from the other end of the performance spectrum. While *4 stars* still is most likely to correspond to a *good* quality hospital, now less than 70% of scores is *4 stars*.

Three-year analysis also shows further improved discrimination (see Figure C 14).

Figure C 13: Scenario 3: Proportion of Superior, Good, and Poor Hospitals by 2-Year Star Scores

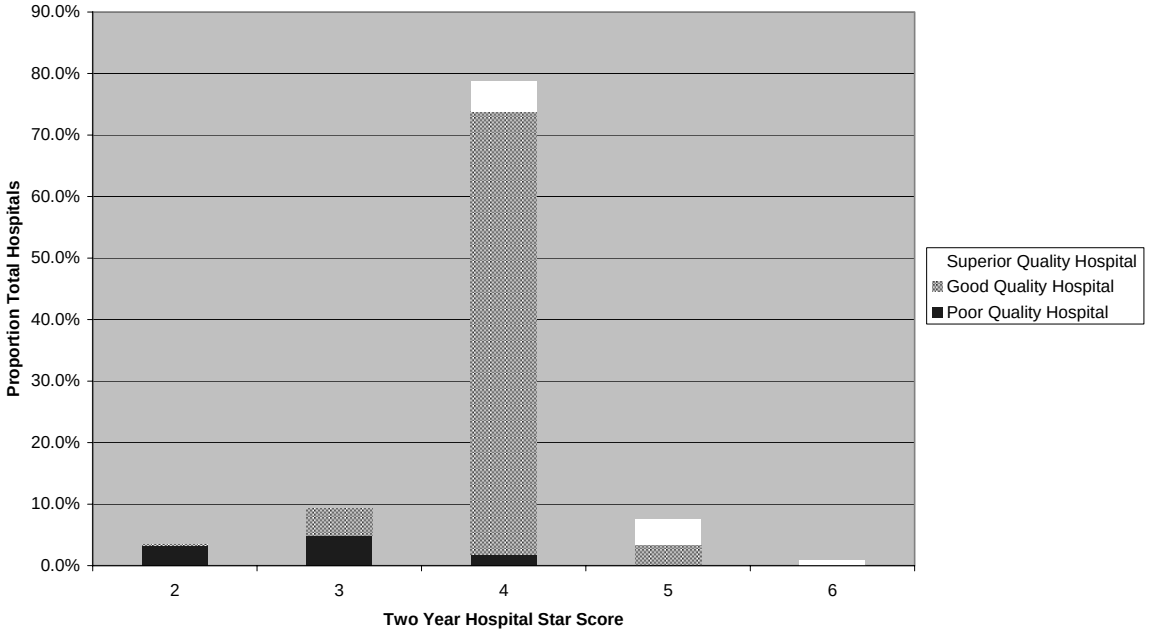
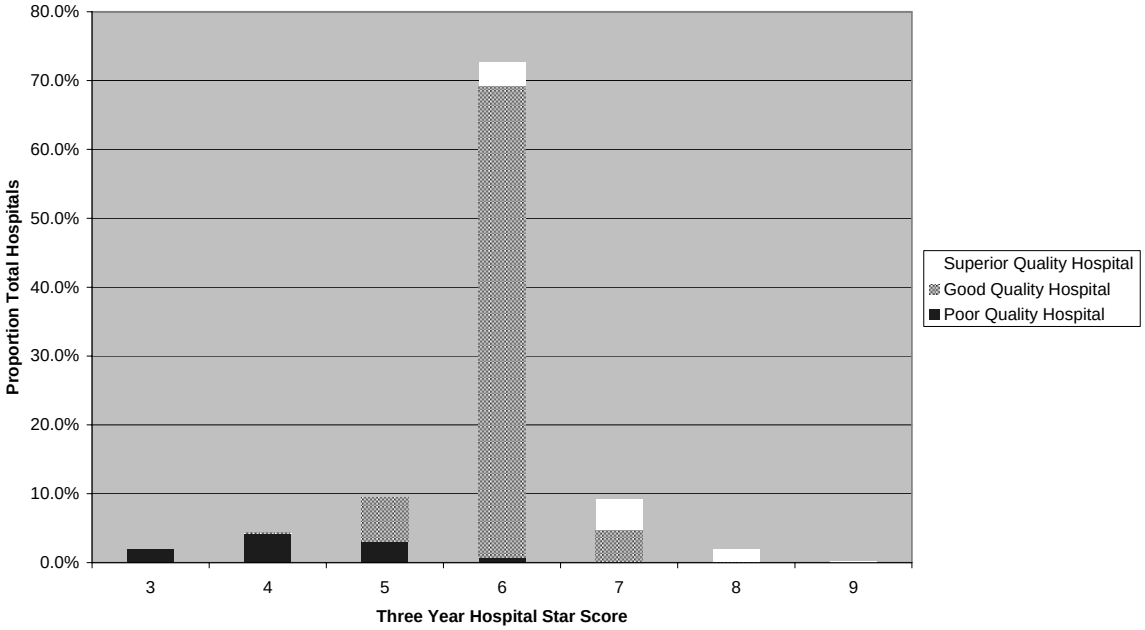
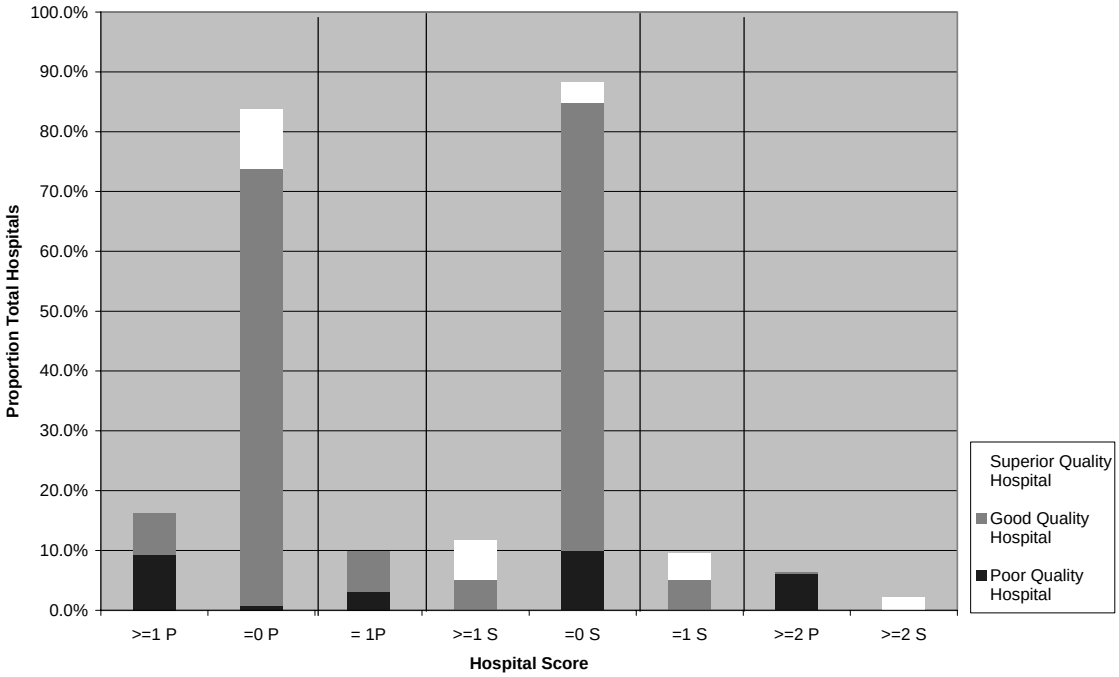


Figure C 14: Scenario 3: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score



Derivative scores also show some promise in this scenario (Figure C 15). There are more hospitals in the very reliably predictive *mostly poor* and *mostly superior* categories.

Figure C 15: Scenario 3: Three-Year Derivative Scores, Predictive Values



For each hospital group, the distribution of scores is summarized in Figure C 16 (which shows the proportion of all hospitals assigned each score, by group) and in Figure C 17 (which shows the proportion of hospitals within each group assigned each score).

Specificity analysis of *ever poor* (Figure C 18) reflects the likelihood that a hospital of either good or superior quality could ever be incorrectly labeled *poor*, even once during the 3-year analysis. *Superior* hospitals are very unlikely to ever receive a *poor* score. *Good* hospitals can infrequently (8.7% of the time) receive one or more *poor* scores (only 0.3% will receive two *poor* scores). *Poor* hospitals almost always (92.5%) receive at least one *poor* score.

Figure C 16: Scenario 3: Expected Distribution of 3-year Star Scores by Hospital Type

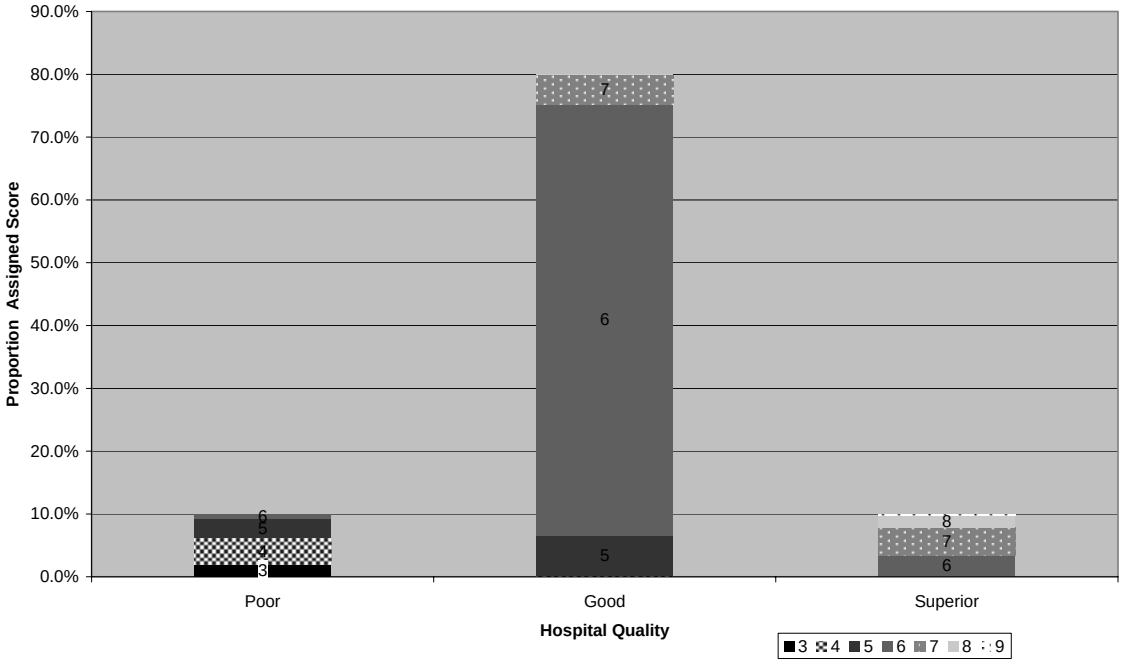


Figure C 17: Scenario 3: Expected Distribution of 3-Year Star Scores by Hospital Type

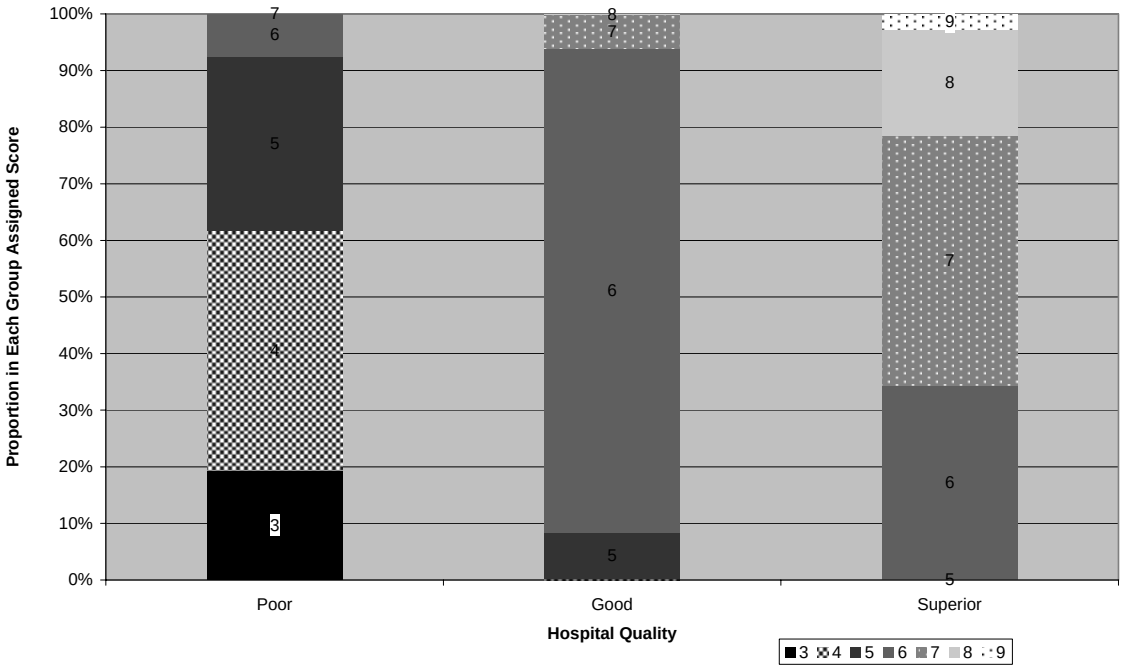
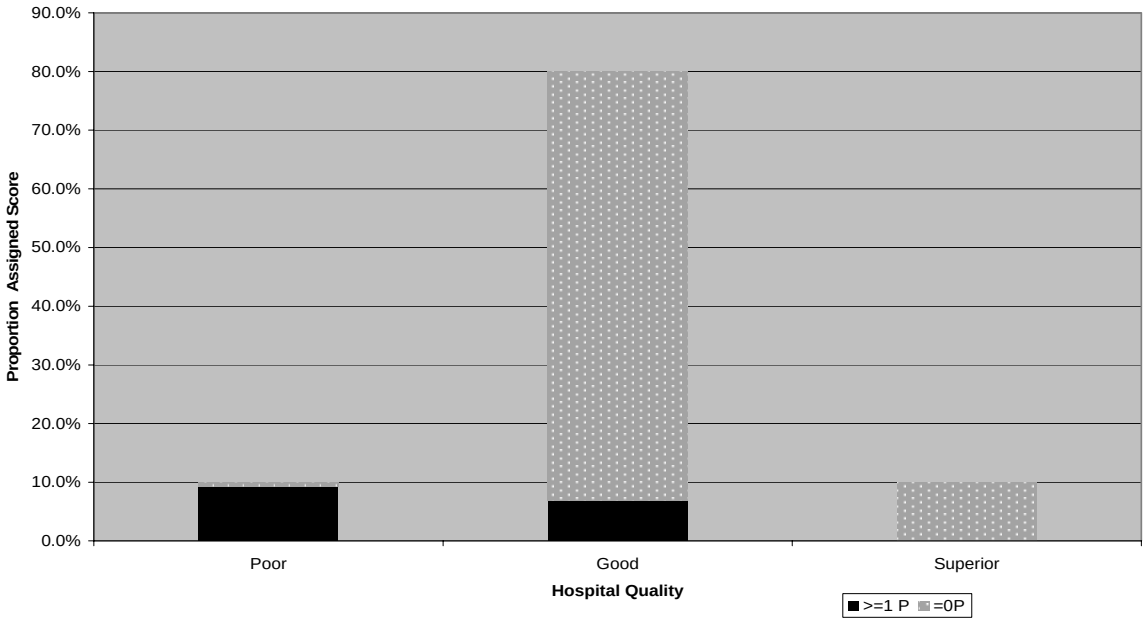


Figure C 18: Scenario 3: In 3 Years Ever Graded Poor vs. Never Graded Poor



Scenario 4: Fewer Patients per Hospital (N = 100)

This scenario explores N: the role of number of patients per hospital. This parameter is part of both the model of the hypothetical hospital world and the evaluation function, in that it is used to calculate the standard deviation for all hospital distributions. Decreasing N makes the distributions of each group wider; the trim points are further out, as seen in Figure C 19.

The results for this scenario (Figure C 20) show that the *star* scores are robust over even fairly small sample sizes

Figure C 19: Scenario 4: Hypothetical World and Evaluation Function

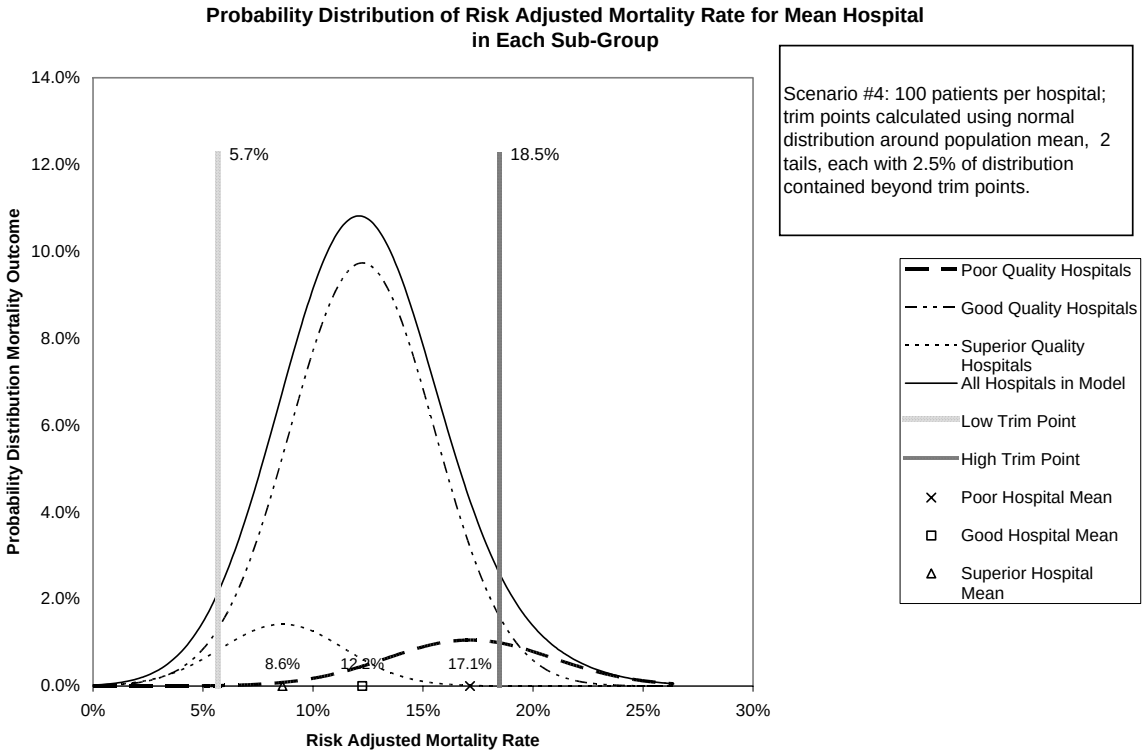
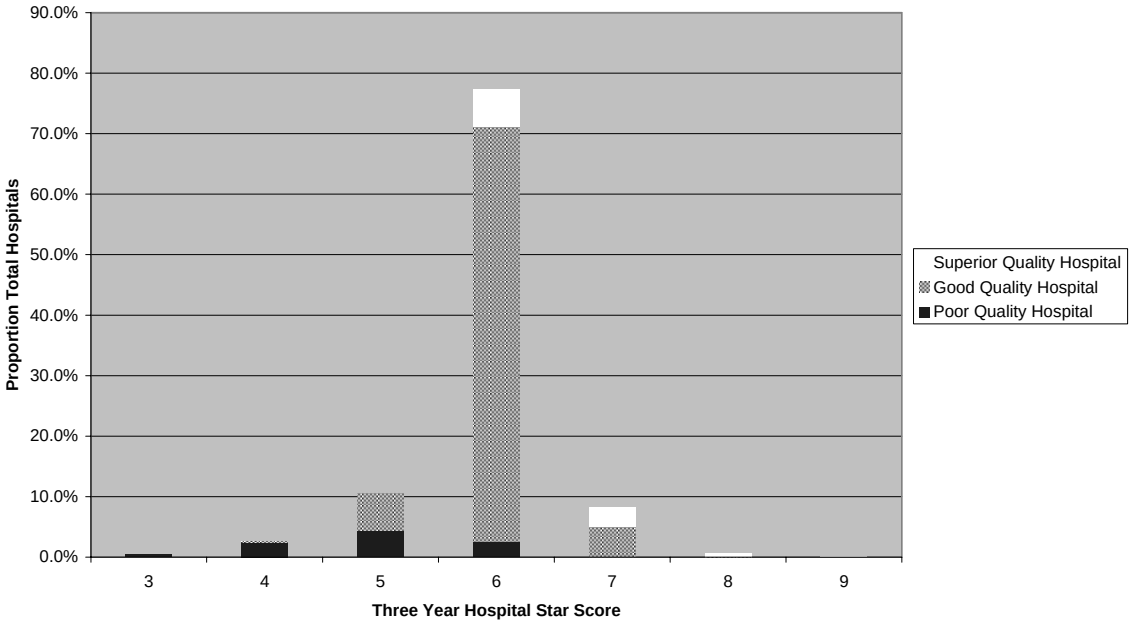
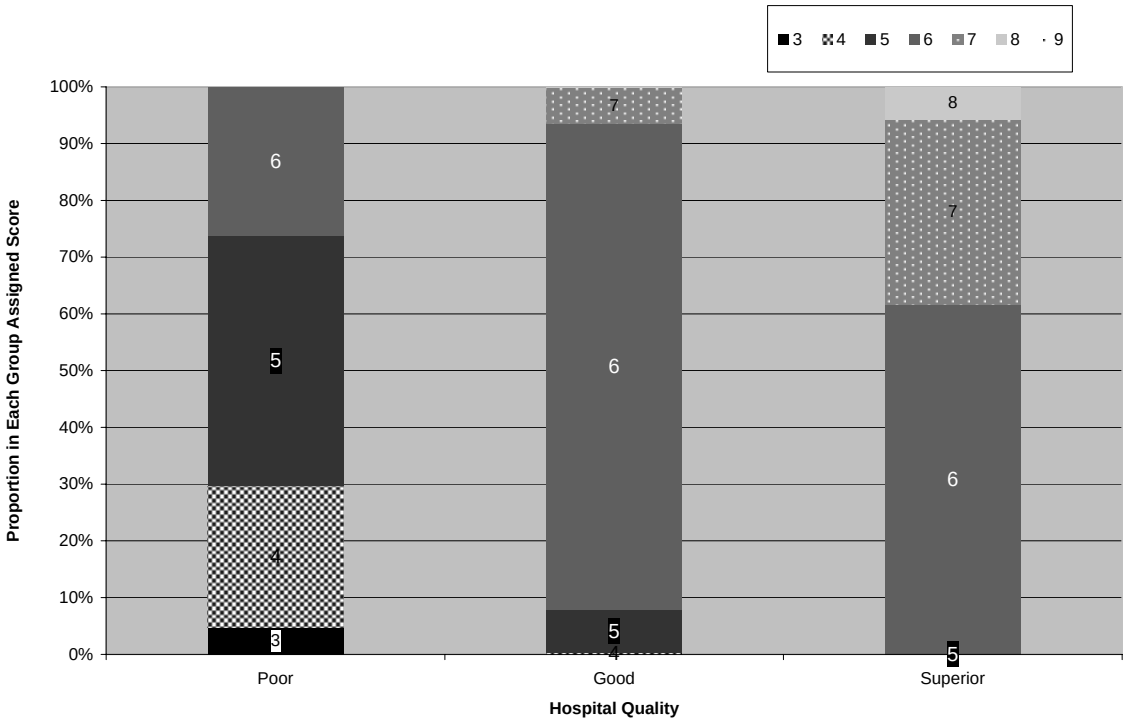


Figure C 20: Scenario 4: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score



Score distributions for each hospital group are summarized in Figure C 21.

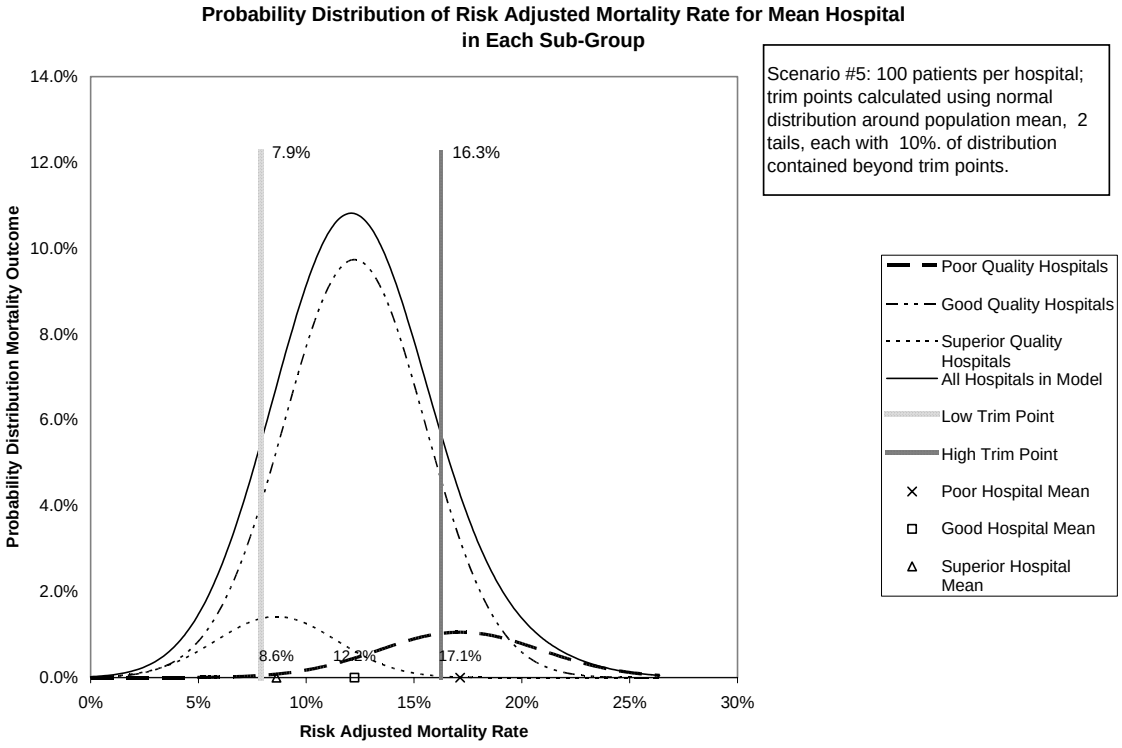
Figure C 21: Scenario 4: Expected Distribution of 3-Year Star Scores by Hospital Type



Scenario 5: Identifying a Higher Proportion of Outliers

In this simulation, the same hypothetical world as in scenario 4 was used; however, the definition of the trim points for the grading function was changed. In this scenario, the trim points are set such that 10% of the overall hospital quality distribution lies to the right of the upper trim point, and 10% lies below the lower trim point (see Figure C 22).

Figure C 22: Scenario 5: Hypothetical World and Evaluation Function



Analysis of scores over 3 years (Figure C 23) shows that by relaxing the trim points, the distribution of scores is spread out as well. There are more hospitals receiving extreme grades. While the more extreme scores are still quite discriminating, there is a very small population of *superior* hospitals which would now receive *5 stars.*

Derivative scores results show (Figure C 24) more hospitals in the useful *mostly poor* and *mostly good* categories. It is still quite rare for a superior hospital to be mislabeled as *poo*r—as evidenced by the *ever poor* (>= 1 P) predictive values.

Figure C 23: Scenario 5: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score

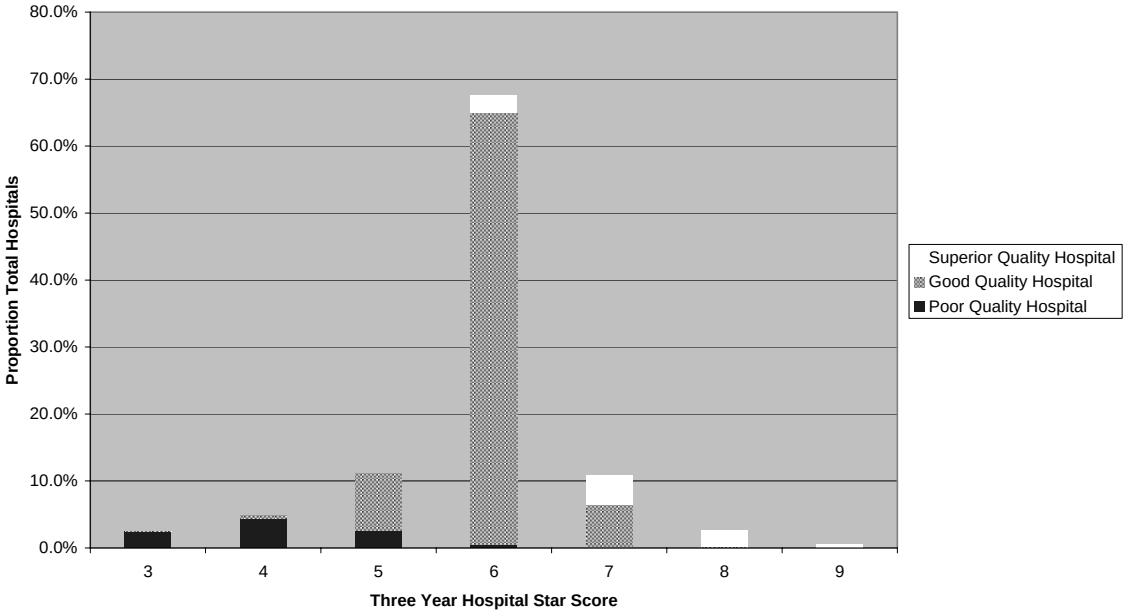
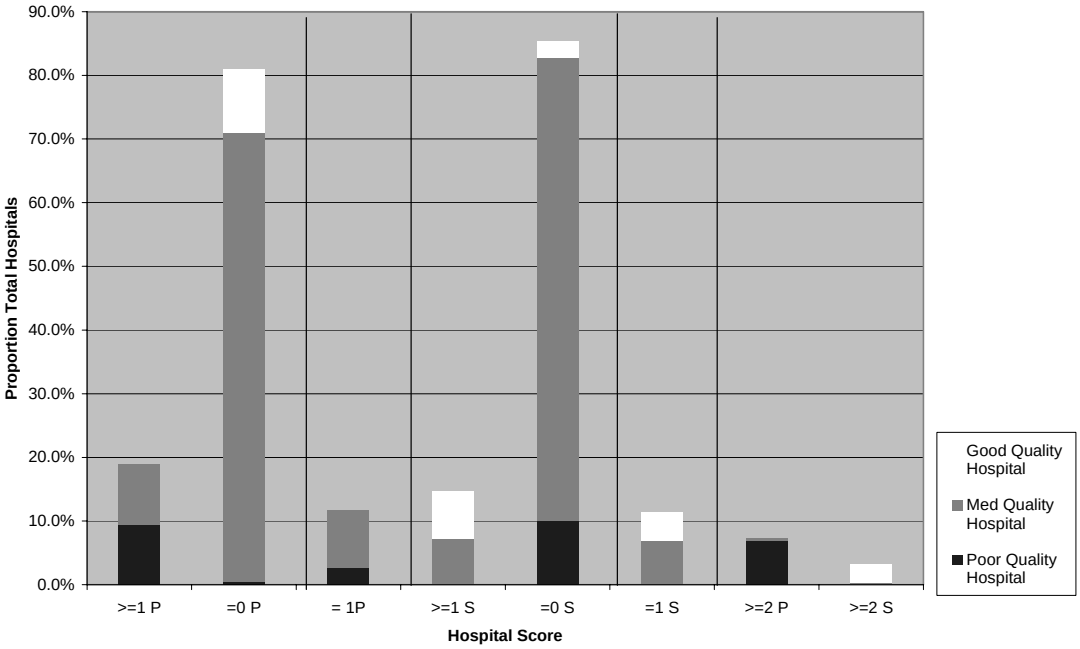
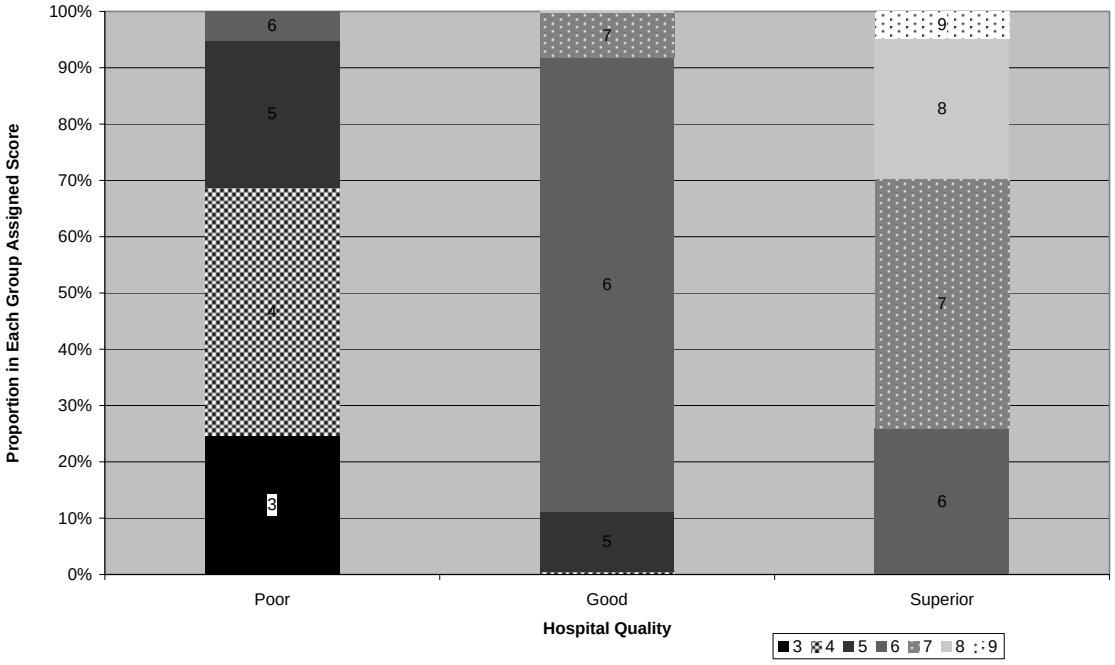


Figure C 24: Scenario 5: Three-Year Derivative Scores, Predictive Values



Scores by hospital group (Figure C 25) confirm this observation. Note that, despite the larger tails there chance that *superior* hospitals will have grades less than *6 stars*, or *poor* hospitals will have grades better than *6 stars*, is almost zero. Grades of *3, 4, 5, 7, 8,* and *9 stars* are therefore useful for at least categorizing hospitals as *not poor* or *not superior*.

Figure C 25: Scenario 5: Expected Distribution of 3-Year Star Scores by Hospital Type



Scenario 6: More Patients per Hospital

This scenario is identical to scenario 5, except that the number of patients per hospital is increased to 400. Results for 3-year analyses are shown in Figure C 26, Figure C 27, and Figure C 28. We see that with greater numbers of patients at each hospital, there can be significant improvement in the ability of the evaluation system to discriminate among classes of hospitals. Using reasonable assumptions for differences in risk-adjusted mortality rates, poor hospitals will receive *3* or *4 stars*; superior hospitals *7, 8, or 9 stars* (with the vast majority receiving *8 or 9*), and good hospitals receive *4-8 stars*, but the majority are concentrated in *5-7 stars*.

Figure C 26: Scenario 6: Proportion of Superior, Good, and Poor Hospital by 3-Year Star Score

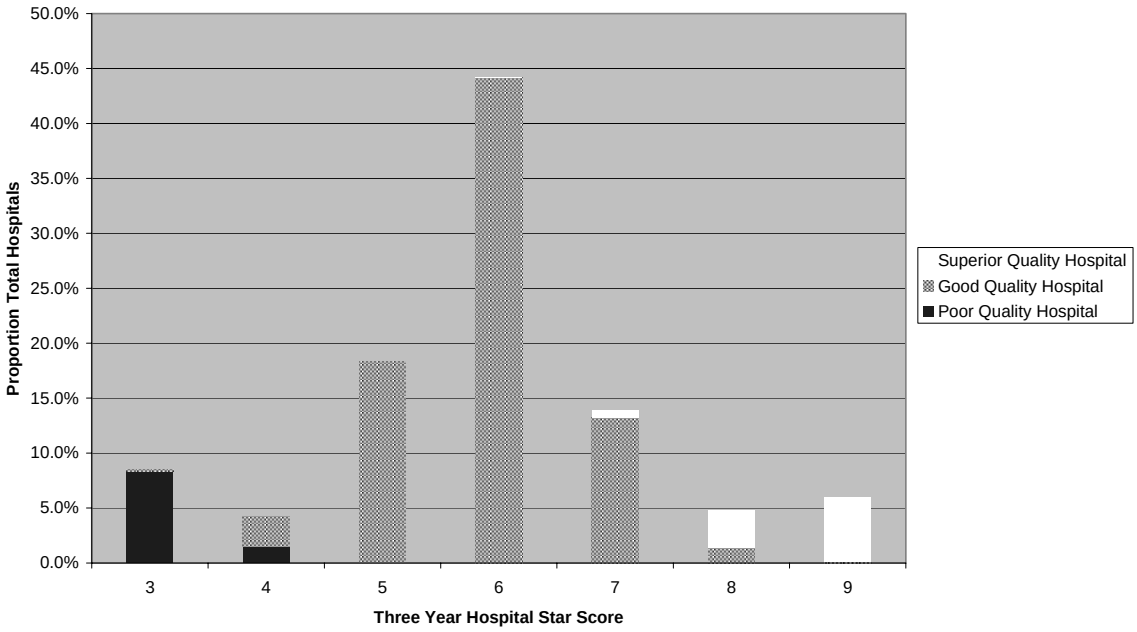


Figure C 27: Scenario 6: Three-Year Derivative Score Predictive Values

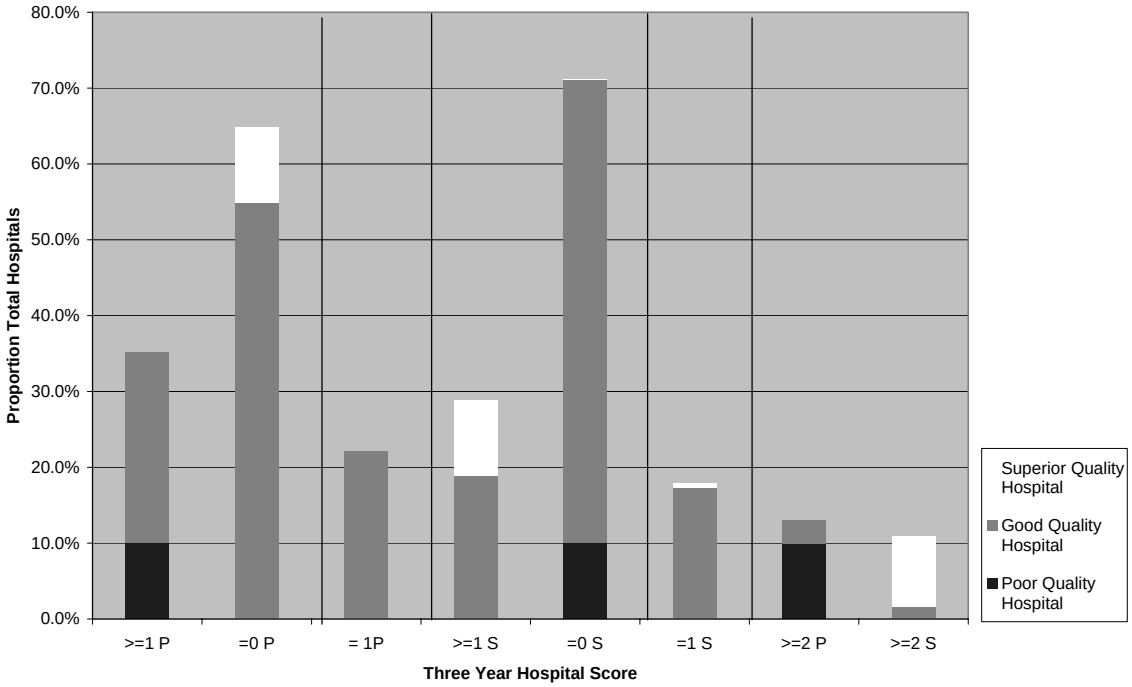


Figure C 28: Scenario 6: Expected Distribution of 3-Year Star Scores by Hospital Type

References

1. Thomas JW, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care.* January 1999;37(1):83-92.

2. Luft HS, Hunt SS. Evaluating Individual Hospital Quality through Outcome Statistics. *JAMA.* May 23/30 1986;255(20):2780-2784.

3. Healthcare Quality and Analysis Division. *Report on Heart Attack Outcomes in California 1996-1998. Volume 1: User's Guide.* Sacramento: California Office of Statewide Health Planning and Development; 2002.

4. Healthcare Quality and Analysis Division. *Report on Heart Attack Outcomes in California 1996-1998. Volume 3: Detailed Statistical Results.* Sacramento: California Office of Statewide Health Planning and Development; 2002.

5. Romano PS, Luft HS, Remy L. *Second Report of the California Hospital Outcomes Project on Acute Myocardial Infarction. Volume Two: Technical Appendix.* Sacramento, CA: California Office of Statewide Health Planning and Development; May 1996.