5. Results of Simulations To Assess the Usefulness of Outcomes Reports

Scenario 1: Reproducing Thomas and Hofer

In this chapter, we will describe the key findings from our simulations. (See Appendix C, available at www.ahrq.gov/clinic/epcindex.htm, for a fuller description of all the results from all of the simulations.)

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 The probability of death at *poor* and *good* hospitals was calculated as in their model as described in an unpublished appendix to their paper.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 5 and Figure 6 above, and Table 14 and Table 15, below.

Notice that in this scenario, a fairly large part of the *poor* quality hospital distribution is intersected by the trim point (Figure 6). 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 14. 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 14: 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 (the probability that a hospital that is actually *poor* will be labeled *poor*) and specificity (the probability that a hospital that is actually *good* will be labeled *good*). 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. Table 15 shows sensitivity and specificity for scenario 1.

Table 15: 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.

Following is a discussion of assessing the evaluation system over multiple years of use.

The results for calculating *star* scores for 2 years are shown in Table 16 and Table 17. 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 16: 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 17: Scenario 1: Expected score distribution over 2 years

	Probability (%) hospital will receive score of--			
What 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 7 and by hospital group in Table 18. 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.

Figure 7: Scenario 1: Percentage of good vs. bad hospitals by 3-year star score

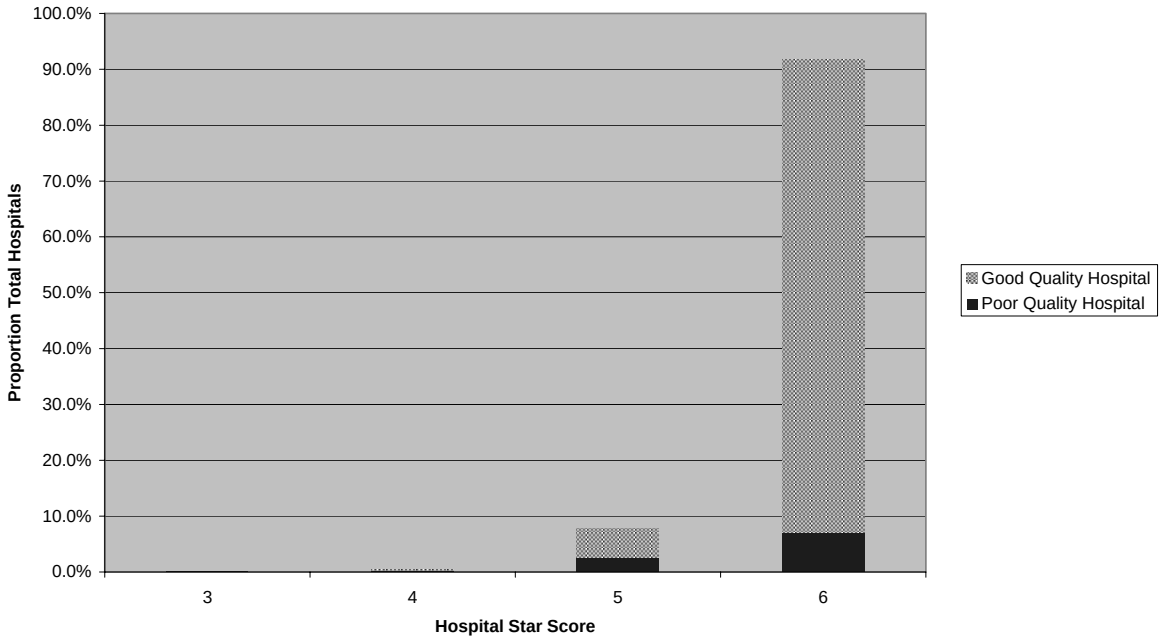


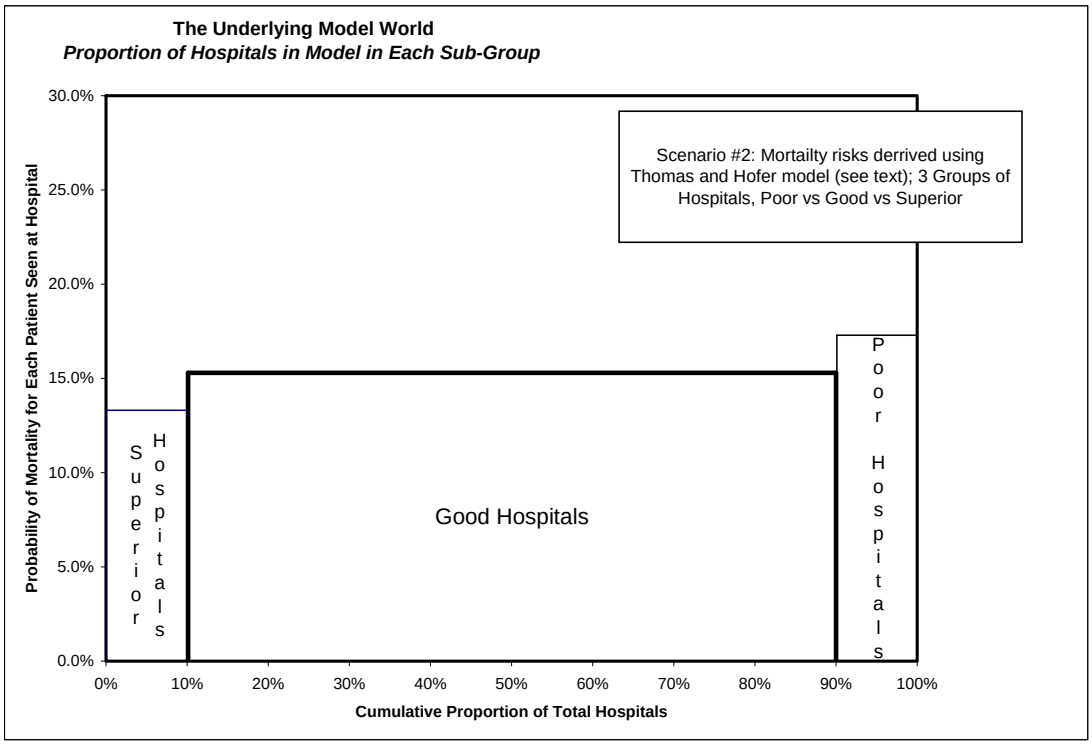
Table 18: Scenario 1: Expected score distribution for good vs. poor hospitals over 3 years

What hospital *really* is	Probability (%) 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%

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 (that is, mortality rates were assumed to be 13.3%, 15.3%, and 17.3% for *superior*, *good*, and *poor* hospitals, respectively, Figure 8). This assumption about *superior* hospitals is arbitrary and meant simply to be approximately as conservative Thomas and Hofer's original assumptions.

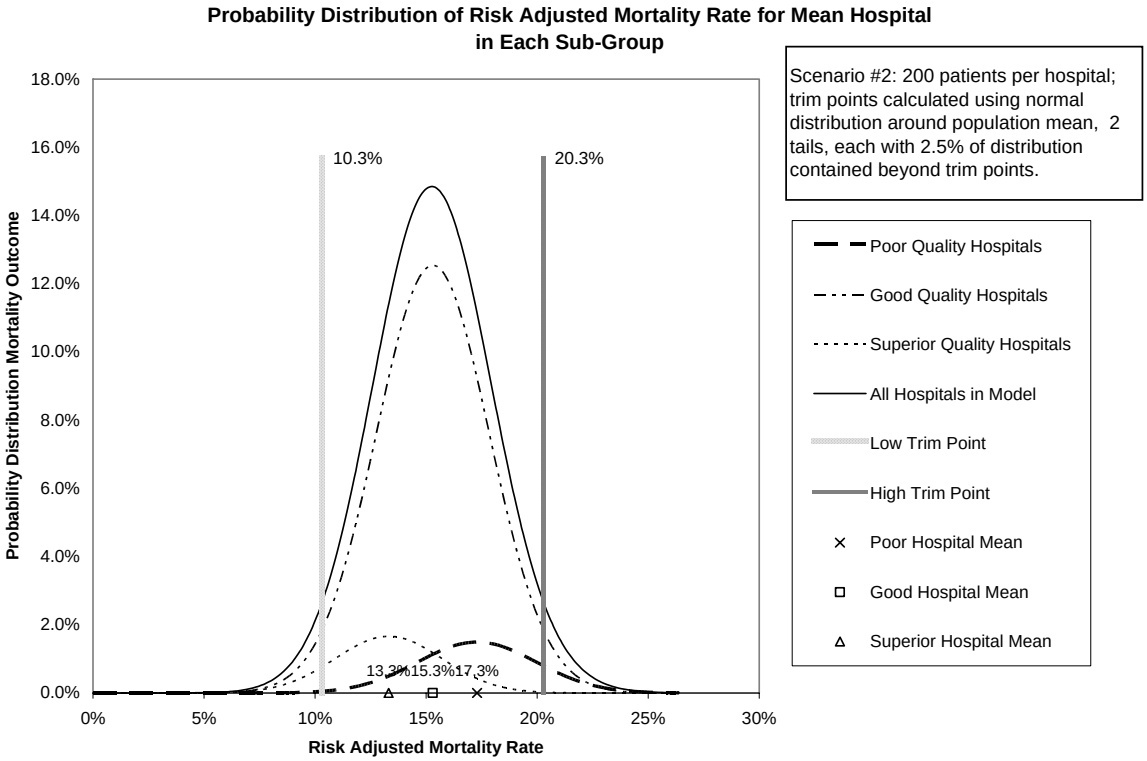
Figure 8: Scenario 2: Hypothetical world of hospitals



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 9.

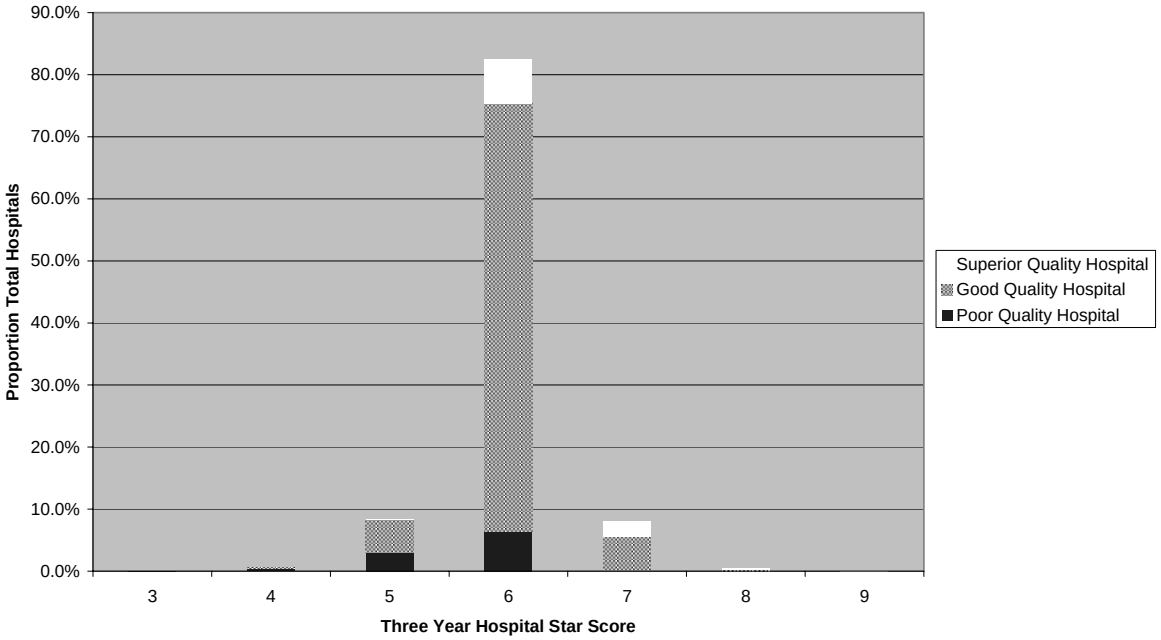
Figure 9: Scenario 2: Hypothetical world and evaluation function



Since there are three possible labels hospitals could receive, simulation 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).

Three-year *star* scores now reliably identify a handful of hospitals at the extremes of mortality scores (Figure 10). 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 10: Scenario 2: Proportion of superior, good, and poor hospitals by 3-year star score

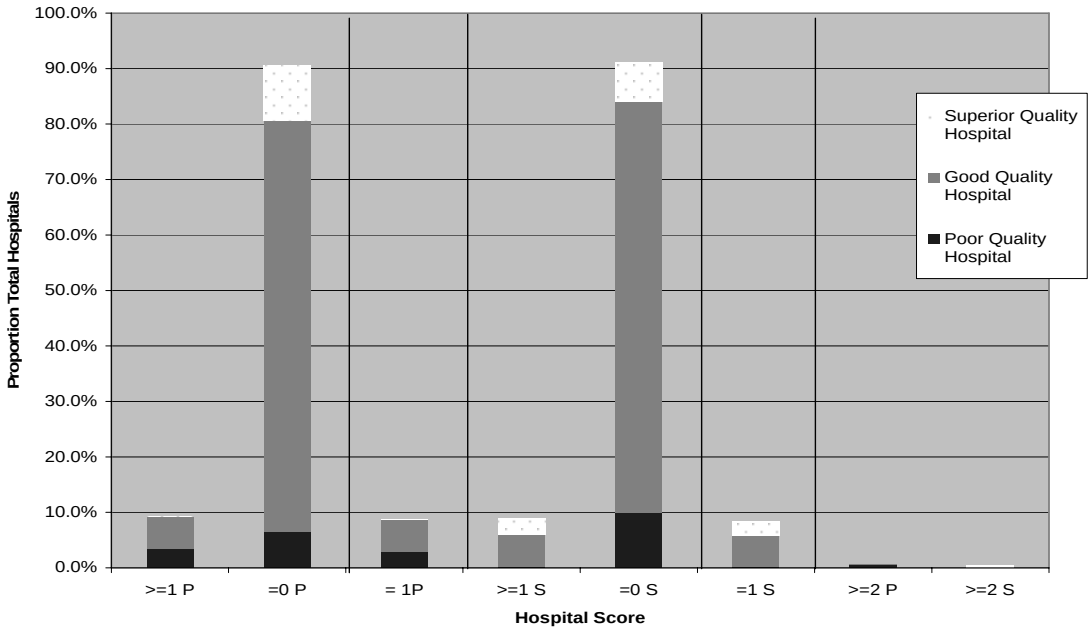


Derivative scores were used to assess whether further discrimination could be obtained among the three sub-groups. The measures are *never poor* (= 0 P), *ever poor* (>= 1 P), *exactly 1 poor* (= 1 P), *mostly poor* (>= 2 P), *never superior* (= 0 S), *ever superior* (>= 1 S), *exactly 1 superior* (= 1 S), and *mostly superior* (>= 2S). The derivative scores for scenario 2 are shown in Figure 11.

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.

Analysis of scenario 2 demonstrated that there could be some improvements to the labels generated by the evaluation system through the addition of multiple hospital 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.

Figure 11: Scenario 2: Proportion of poor, good, and superior hospitals with each type of derivative score



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.[67, 68] (See Appendix B for the algorithm used to generate the mean mortality for each group.)

The model world is shown in Figure 12 and the evaluation function is summarized in Figure 13. The evaluation function is based on the reported population mean mortality rate and 2.5% trim points, as described above.

Figure 12: Scenario 3: The hypothetical world

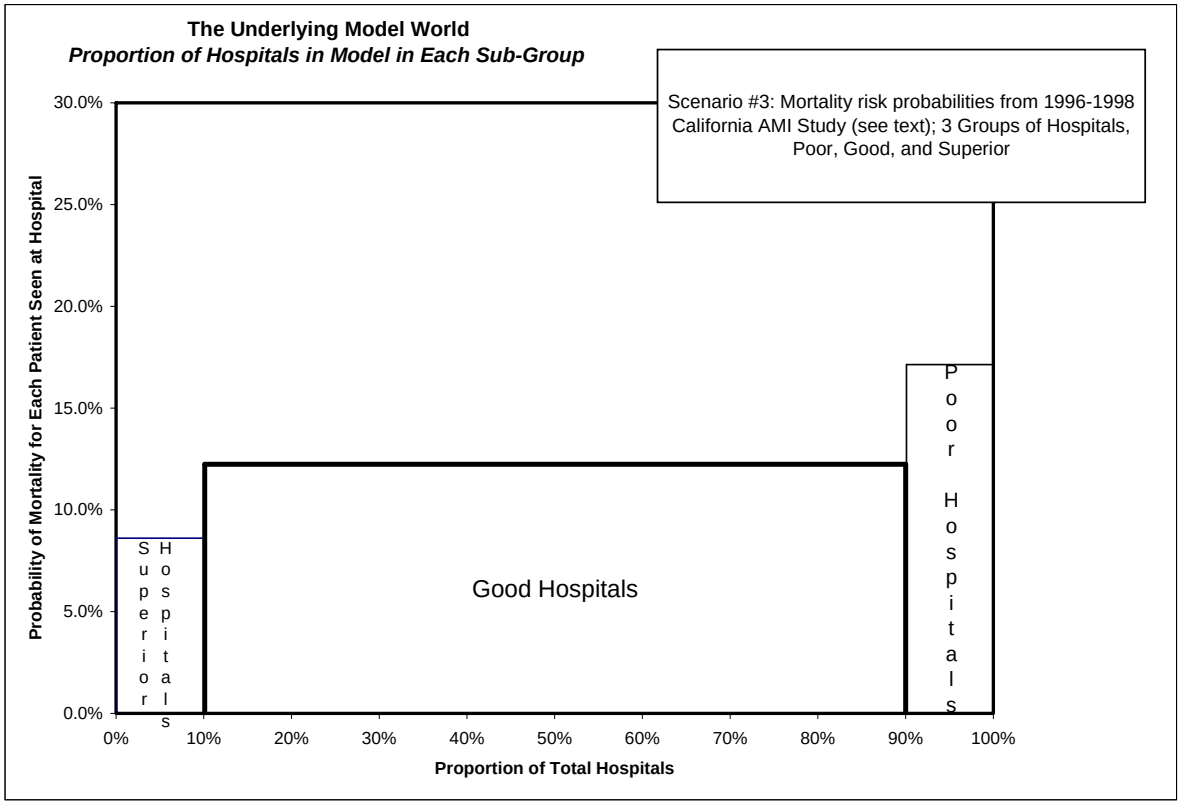
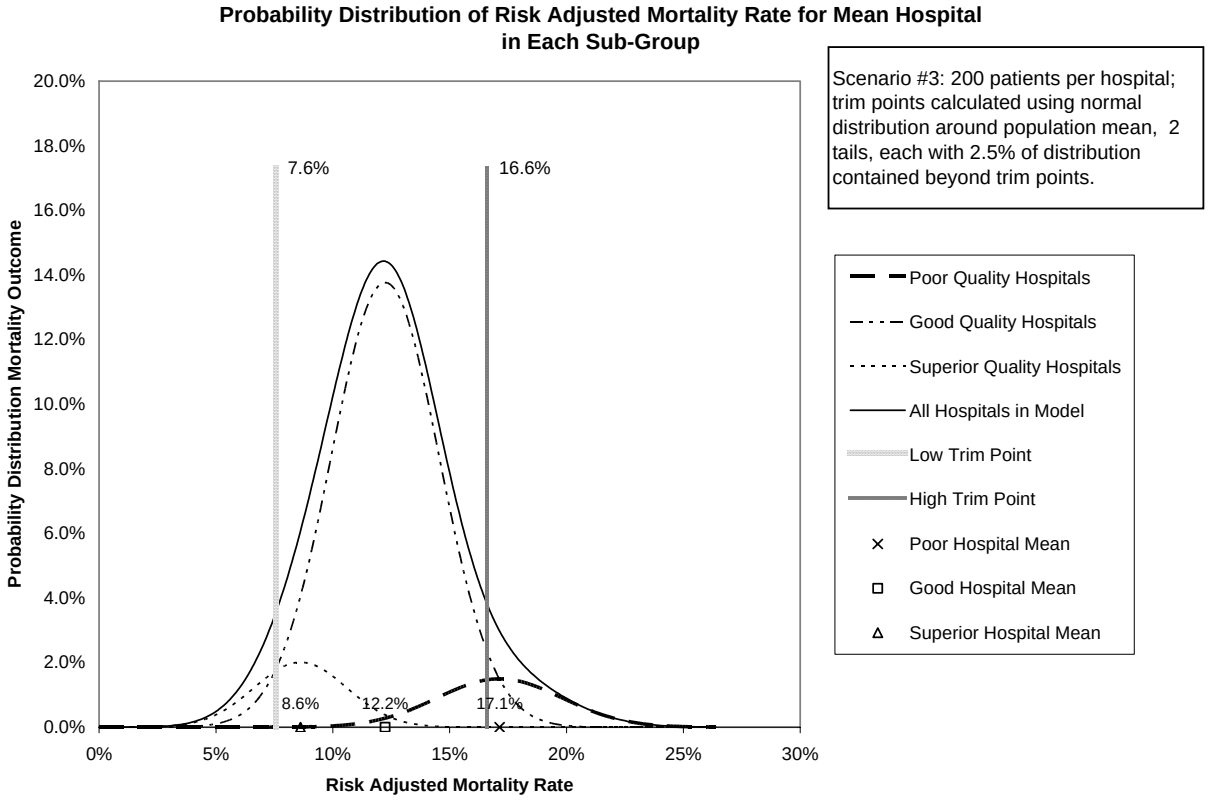
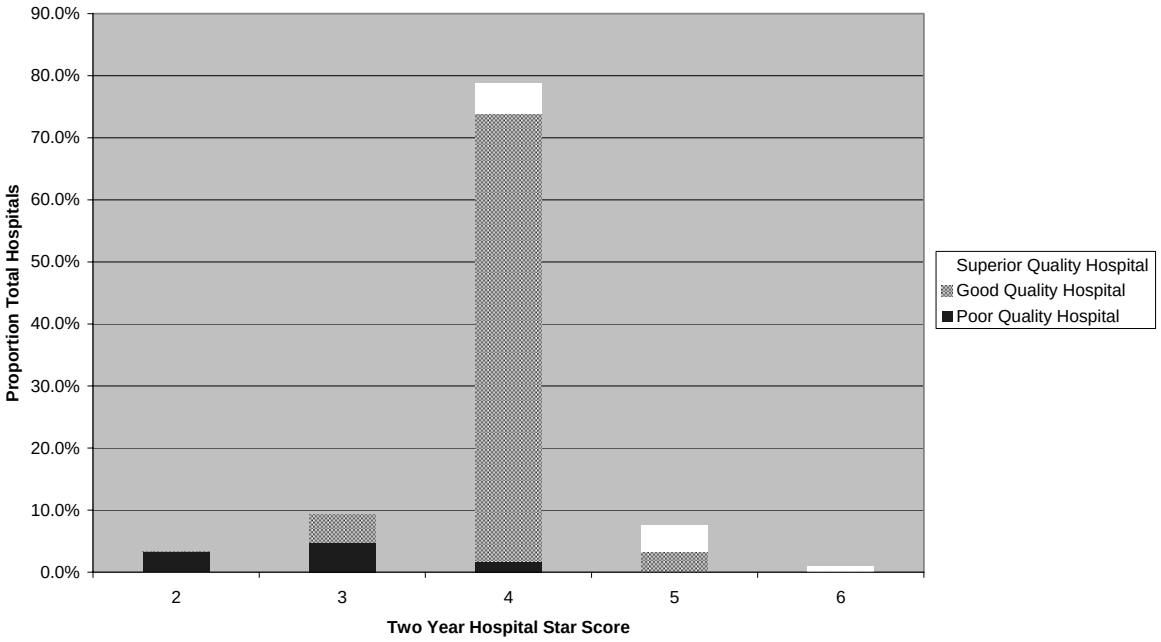


Figure 13: Scenario 3: Hypothetical world and evaluation function



The greater difference between mortality rates in the *superior* and *poor* groups has resulted in better discrimination in even in just 2 years of reporting (see Figure 14). 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*.

Figure 14: Scenario 3: Proportion of superior, good, and poor hospitals by 2-year star scores



Three-year analysis also shows further improved discrimination (see Figure 15). Derivative scores also show some promise in this scenario (Figure 16). There are more hospitals in the very reliably predictive *mostly poor* and *mostly superior* categories. *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.

For each hospital group, the distribution of scores is summarized in Figure 17.

Figure 15: Scenario 3, year 3: Proportion of superior, good, and poor hospitals by 3-year star score

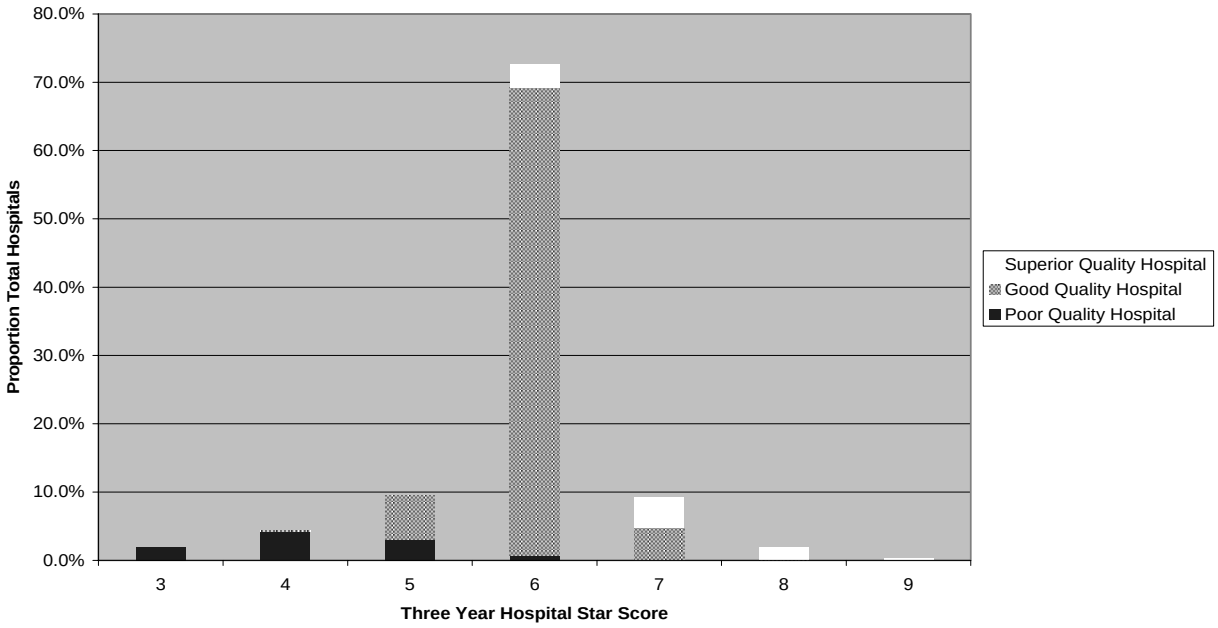


Figure 16: Scenario 3: Three-year derivative scores, predictive values

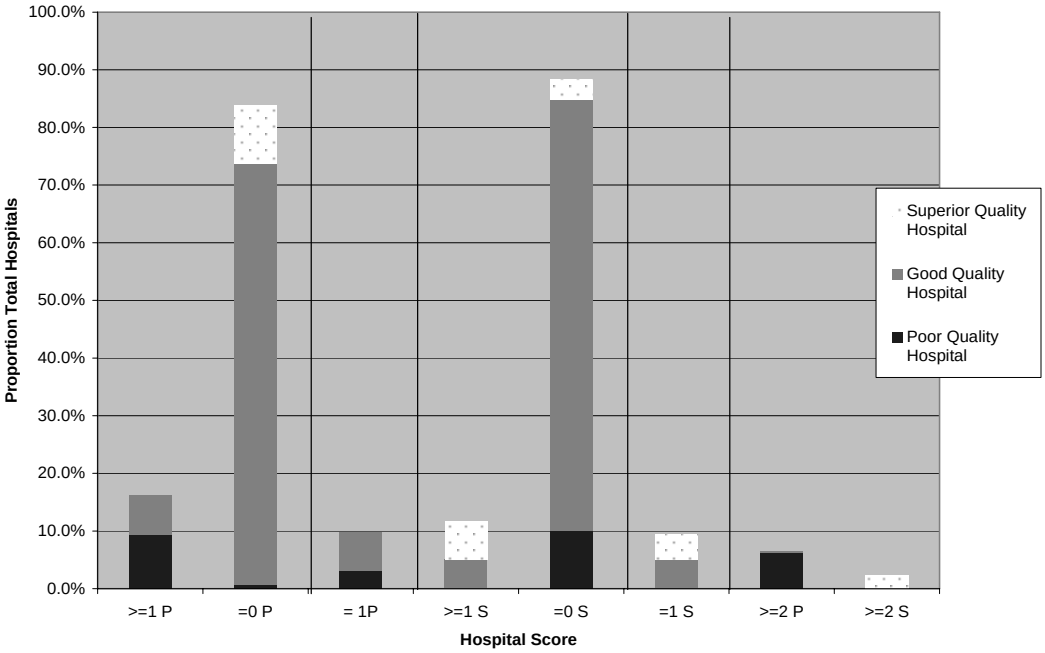
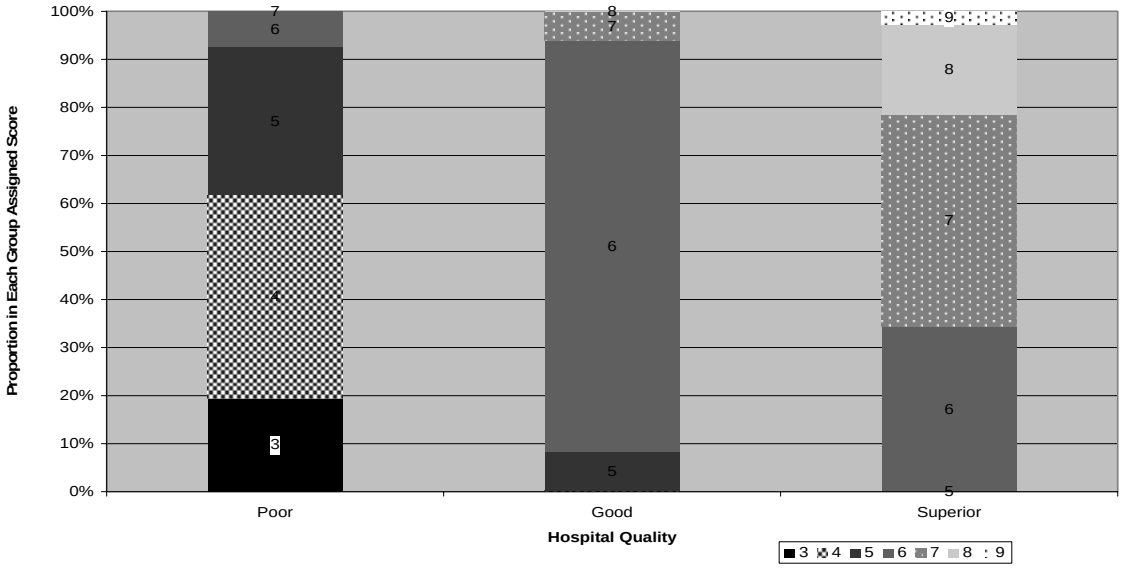


Figure 17: Scenario 3: Distribution of 3-year derivative scores, predictive values



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 18.

The results for this scenario (Figure 19) show that the *star* scores are robust, despite the smaller sample size.

Figure 18: Scenario 4: Hypothetical world and evaluation function

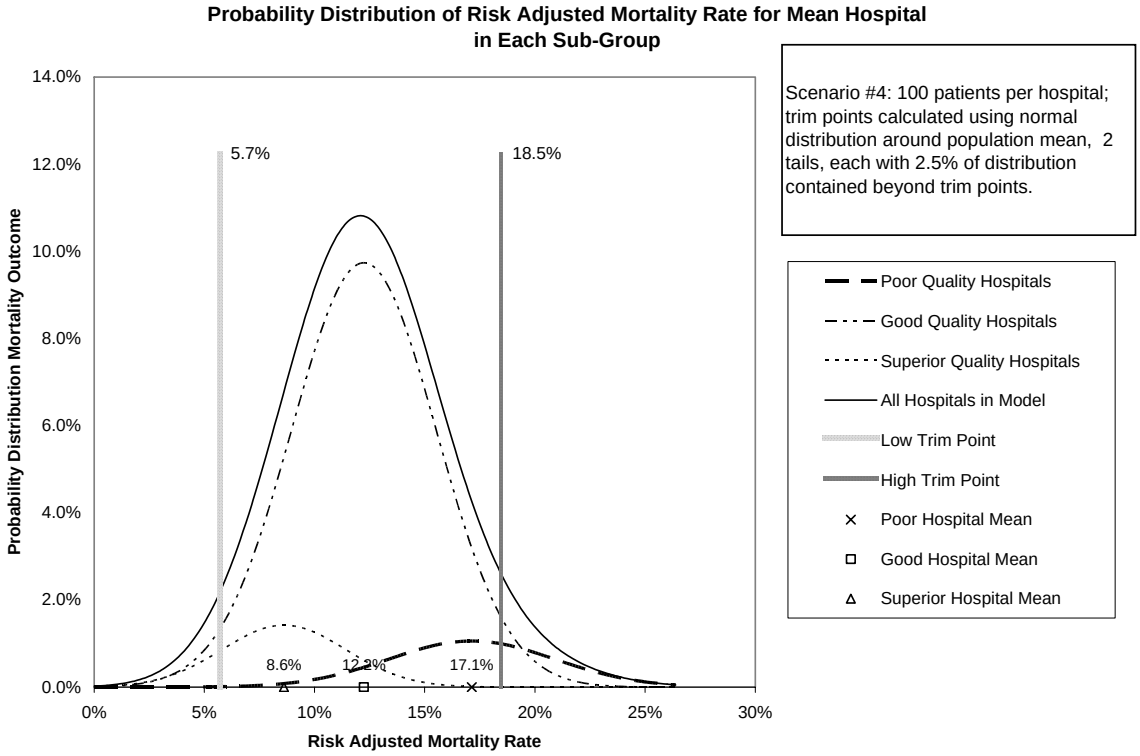
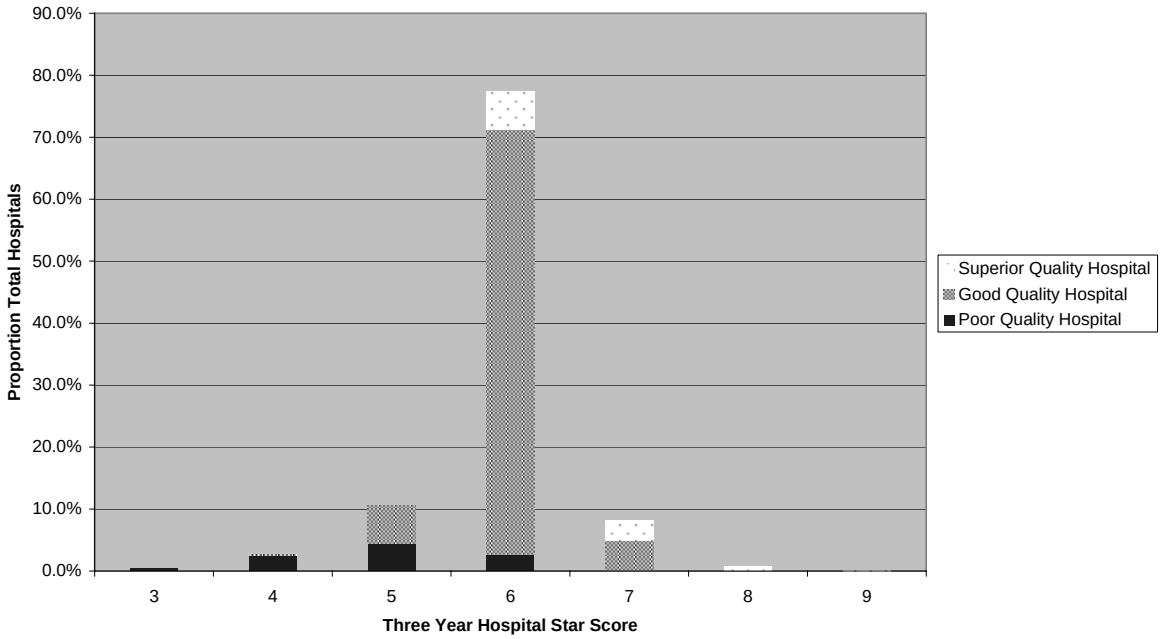


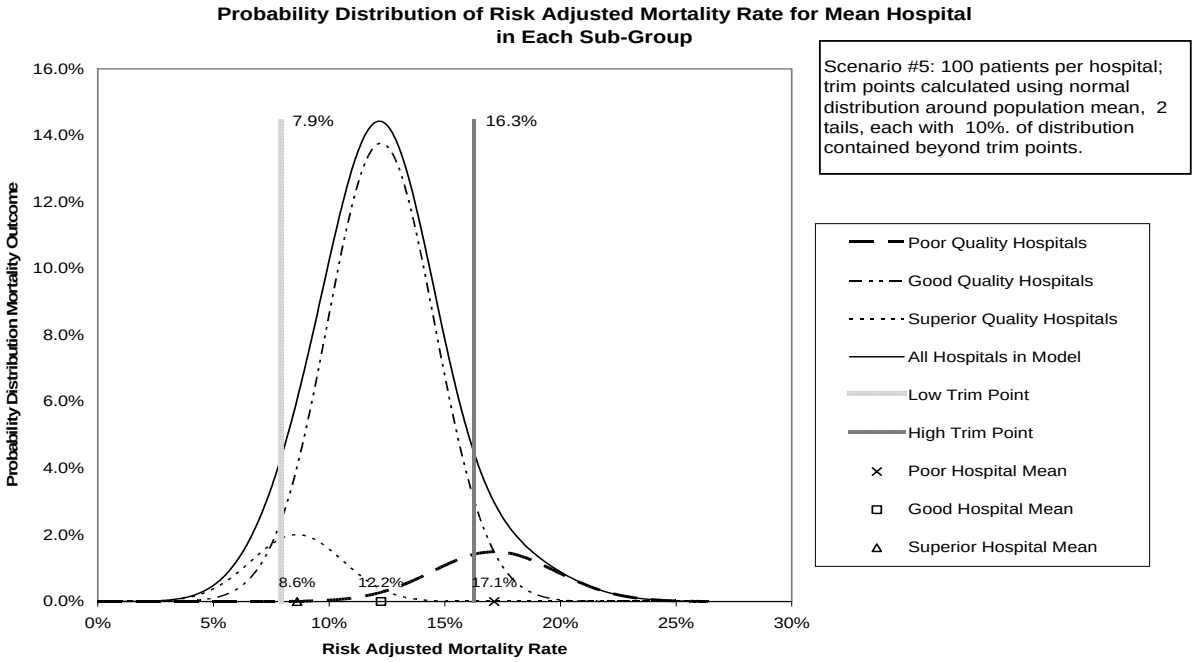
Figure 19: Scenario 4, year 3: Proportion of superior, good, and poor hospitals by 3-year star score



Scenario 5: Identifying a Higher Proportion of Outliers

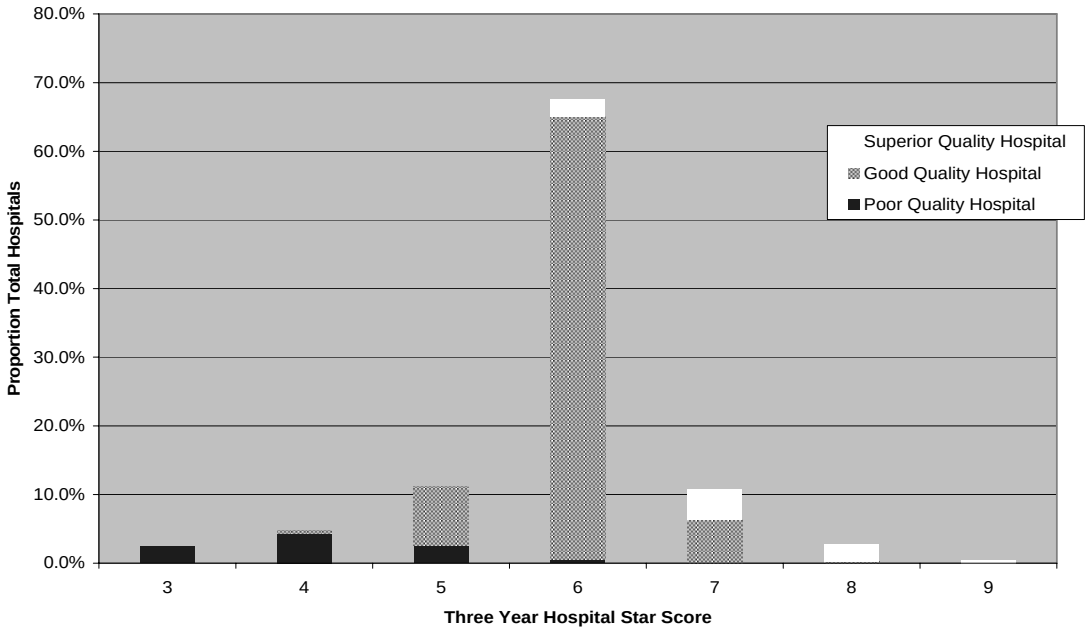
In this simulation, the same hypothetical world as in scenario 3 was used, however, the definition of the trim points for the grading function was changed. In this scenario, the trim points are set such that 10% of the overall hospital quality distribution lies to the right of the upper trim point, and 10% lies below the lower trim point (see Figure 20).

Figure 20: Scenario 5: Hypothetical world and evaluation function



Analysis of scores over three years (Figure 21) shows that by relaxing the trim points, the distribution of scores is spread out as well. There are more hospitals receiving extreme grades. 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 21: Scenario 5: Proportion of superior, good, and poor hospitals by 3-year star score



Scenario 6: More Patients per Hospital

This scenario is discussed in more detail in Appendix C. When the number of patients per hospital is increased to 400, discrimination by star score or derivative scores becomes very good.