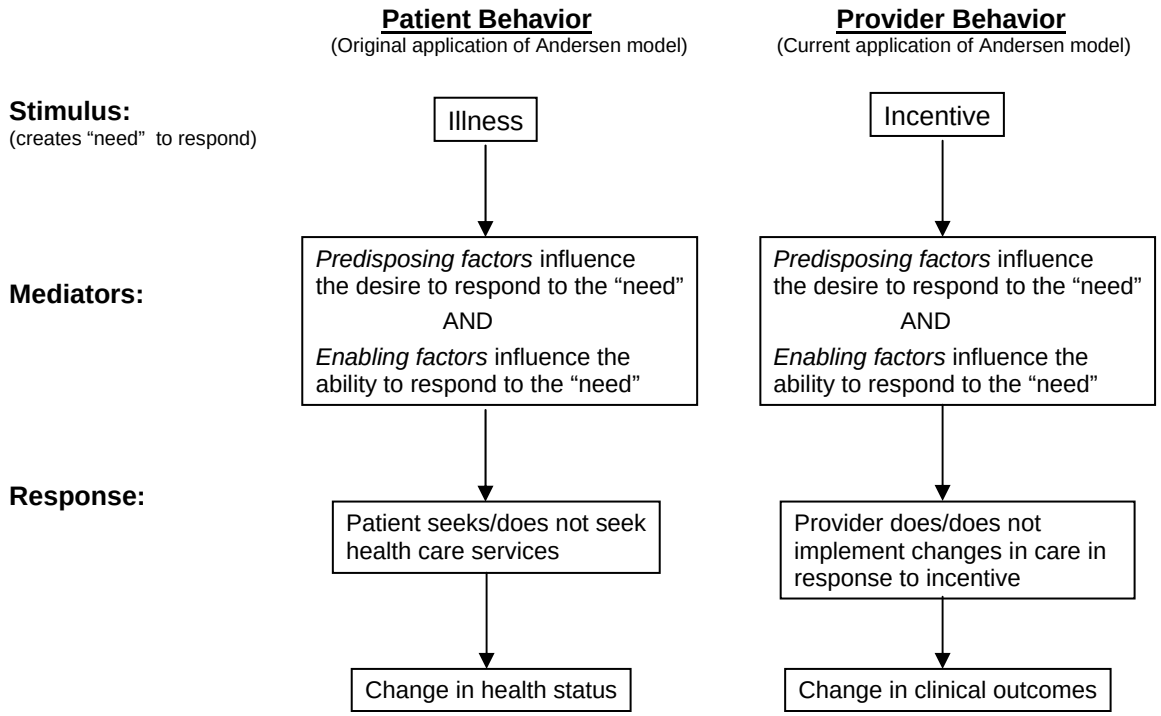
to access care. Similarly, organizational resources (e.g., of the clinic in which the provider practices) could have an enabling effect on provider behavior just as community resources influence patient actions.

Figure 1: Application of Andersen's model to provider behavior

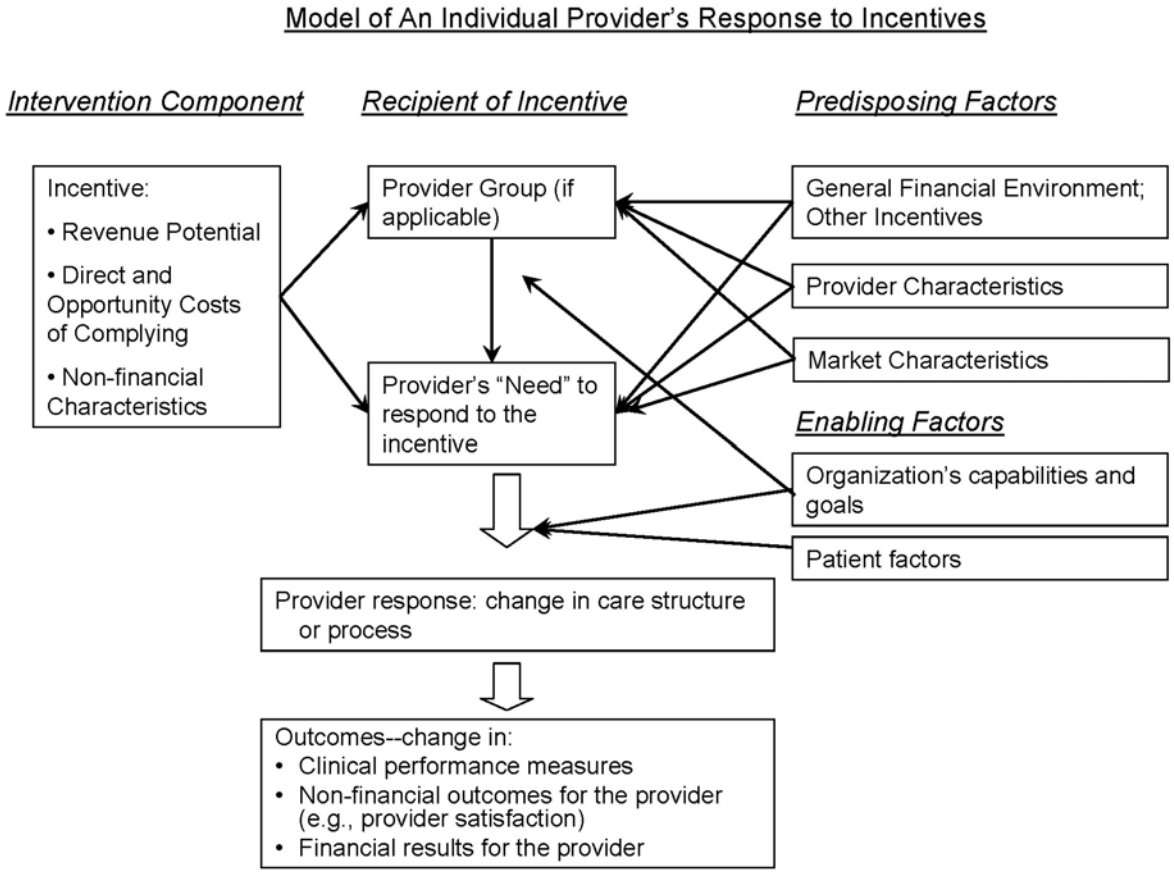


This model complements and integrates, rather than replaces, the extant economic, psychology, and decision and organizational theory literature on incentives. For instance, principal-agent theory from economics is useful for assessing the tradeoffs between different incentive structures and how these might vary as a function of the health plan's ability to mandate provider behavior or monitor different aspects of provider performance.[21, 22, 26] Principal-agent models emphasize the risk to the plan that a provider might shirk or provide poor quality. Similarly, reinforcement theorists have pointed out the potential impact of a variety of types of reinforcers on behavior, including professional and social reinforcement in addition to economic factors.[27] In an excellent review of the economic and psychological theories of incentives, however, Town et al. point out that the potential for bad provider behavior implied in principal-agent analyses and the need for reinforcement implied in reinforcement theory may be countered by strong psychological forces such as expected regret or chagrin if patients have poor outcomes. [26, 28, 29] Frey and Kuhn make analogous points about intrinsic motivation, professionalism, and altruism.[30, 31]

Each of these factors fits into our model, and the model helps explain their relationship to each other. For instance, expected regret about poor performance, intrinsic motivation, and

financial environment and other incentives, as well as by provider characteristics and market variables) and by enabling factors at the organizational and patient levels.

Figure 2: Model of an individual provider's response to incentives



In Figure 3, we show the analog of this model we propose should be used to understand how organizations (i.e., hospitals, medical groups) respond to incentives. This model differs from the model for individual providers in that the charter and mission of an organization are the analog of provider characteristics such as intrinsic motivation and influence the organization's predisposition to respond. Furthermore, congruence with organizational goals is no longer an enabling factor, but goal congruence with individual providers or staff is (see Figure 3).

More research will be needed to assess our labeling of factors as "predisposing" or "enabling", and some factors may both predispose and enable. Fortunately, it is not nearly as important to get the labels correct as to identify potential determinants of behavior so that they can be explicitly studied.

Figure 3: Model of an organization's response to incentives

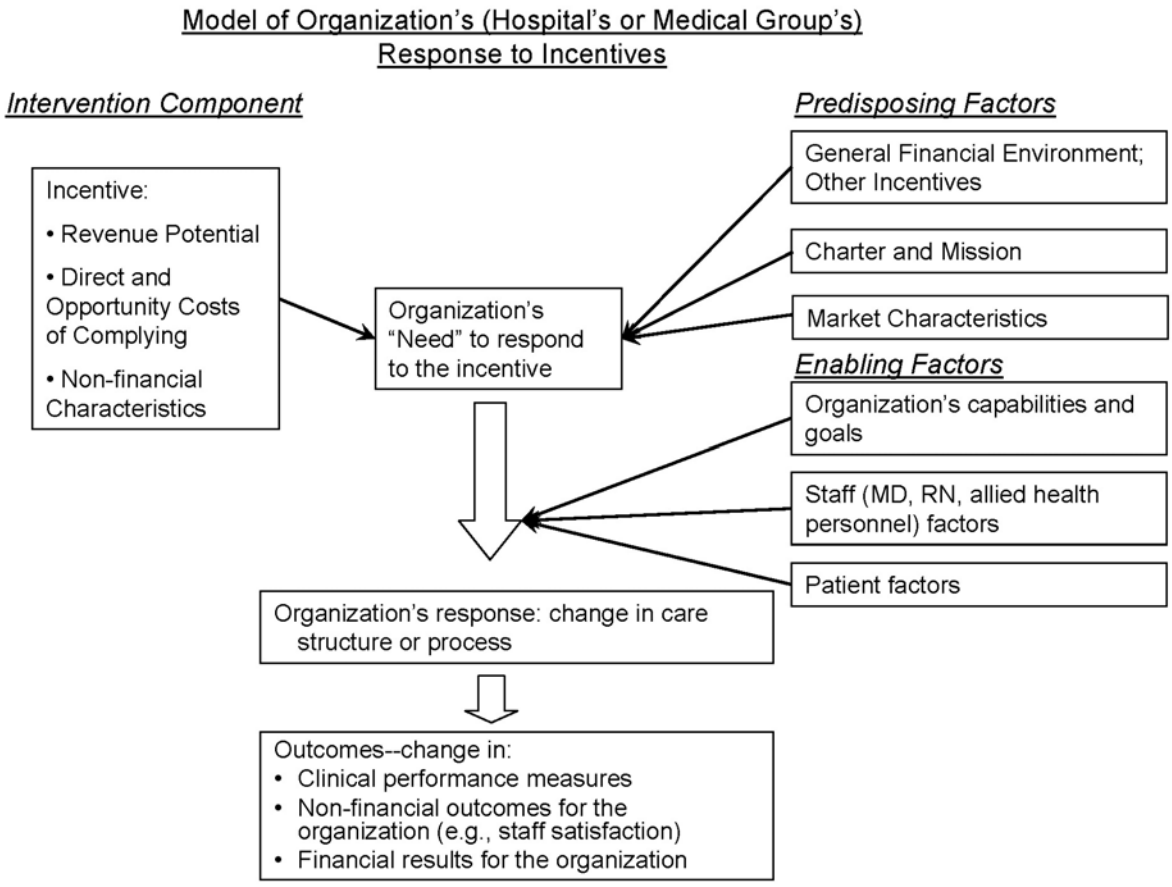


Figure 4: Articles Identified By Systematic Searches

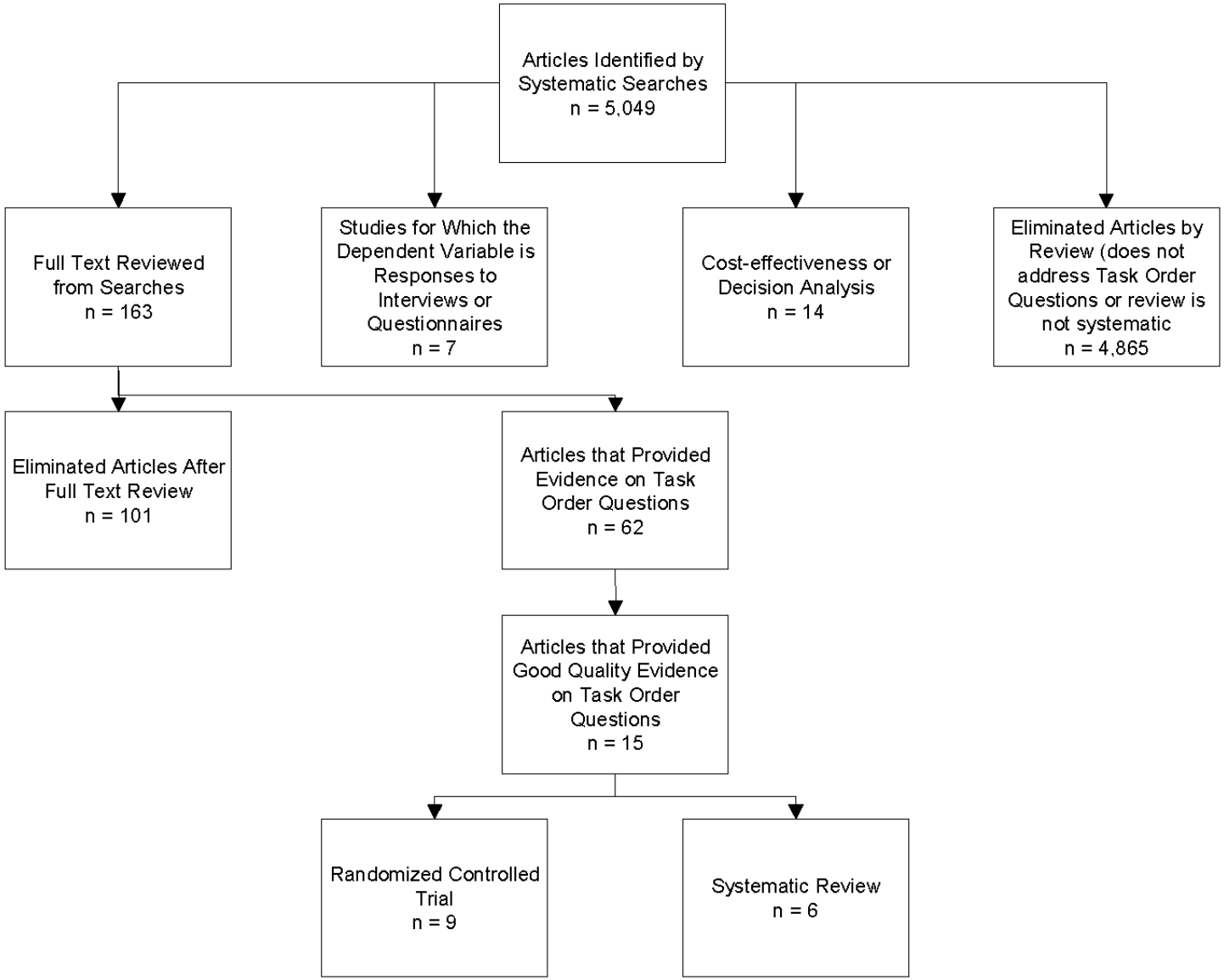
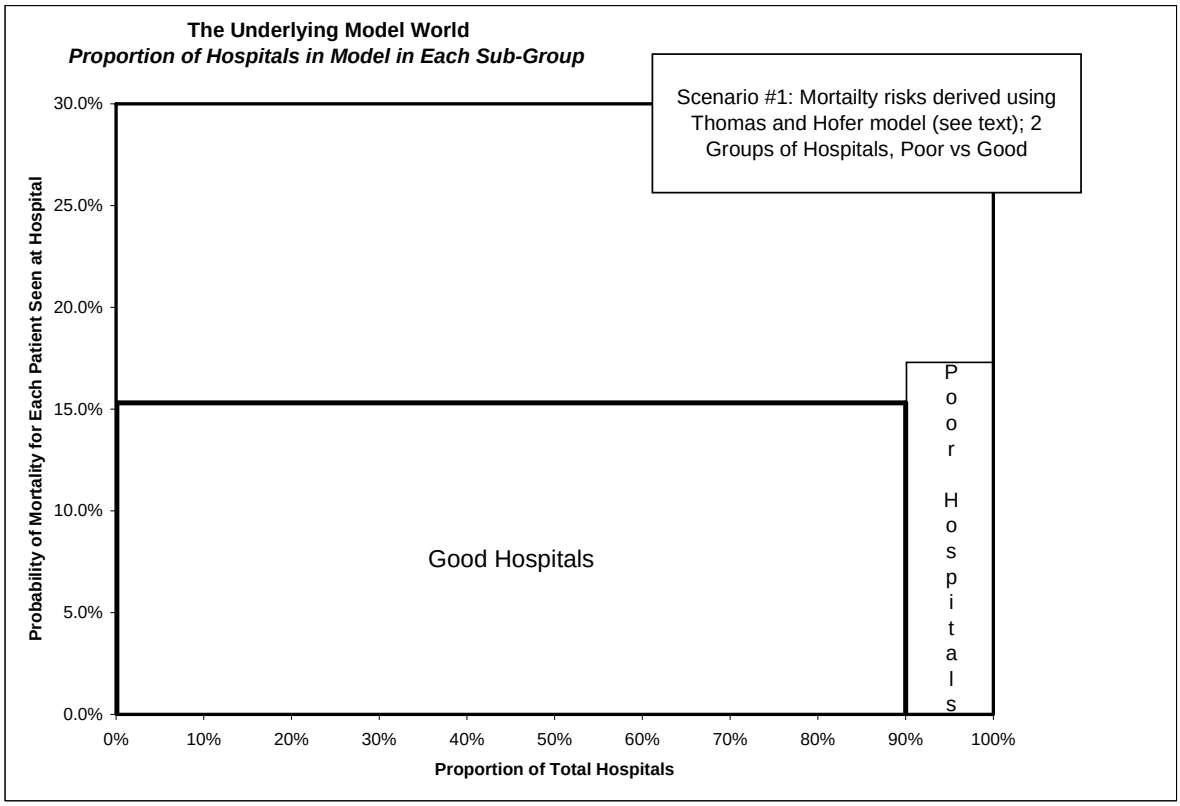
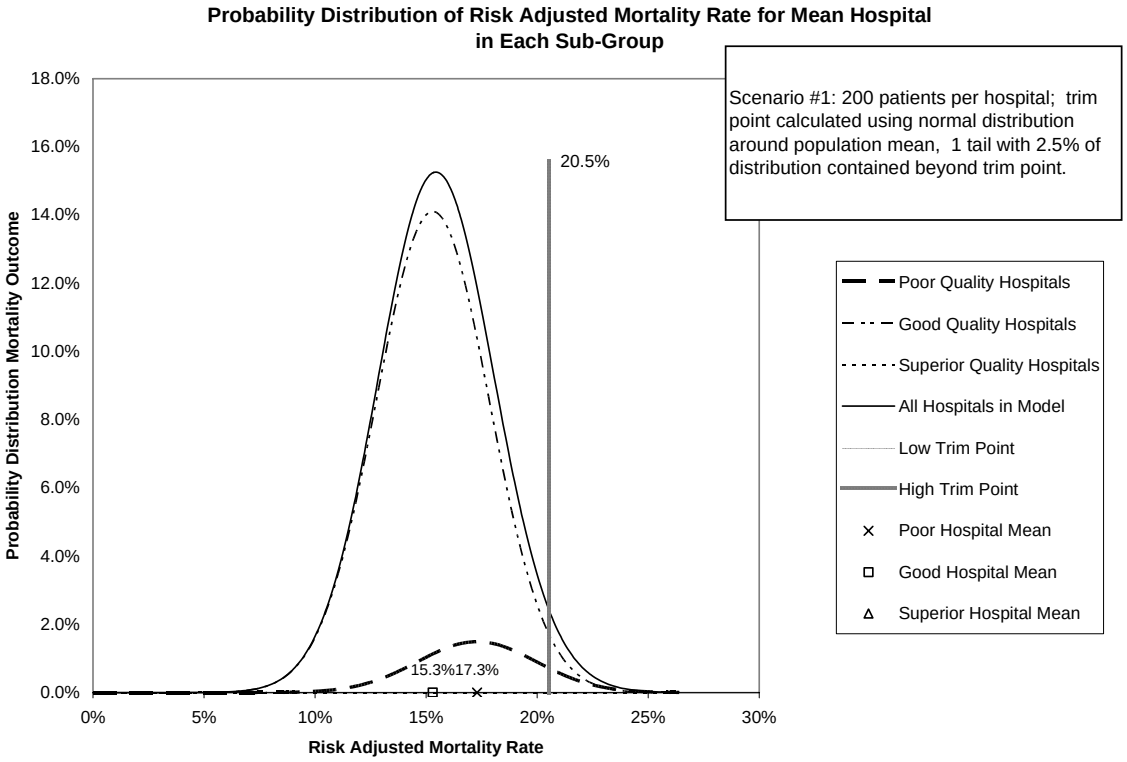


Figure 5: Hypothetical world of hospitals



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

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

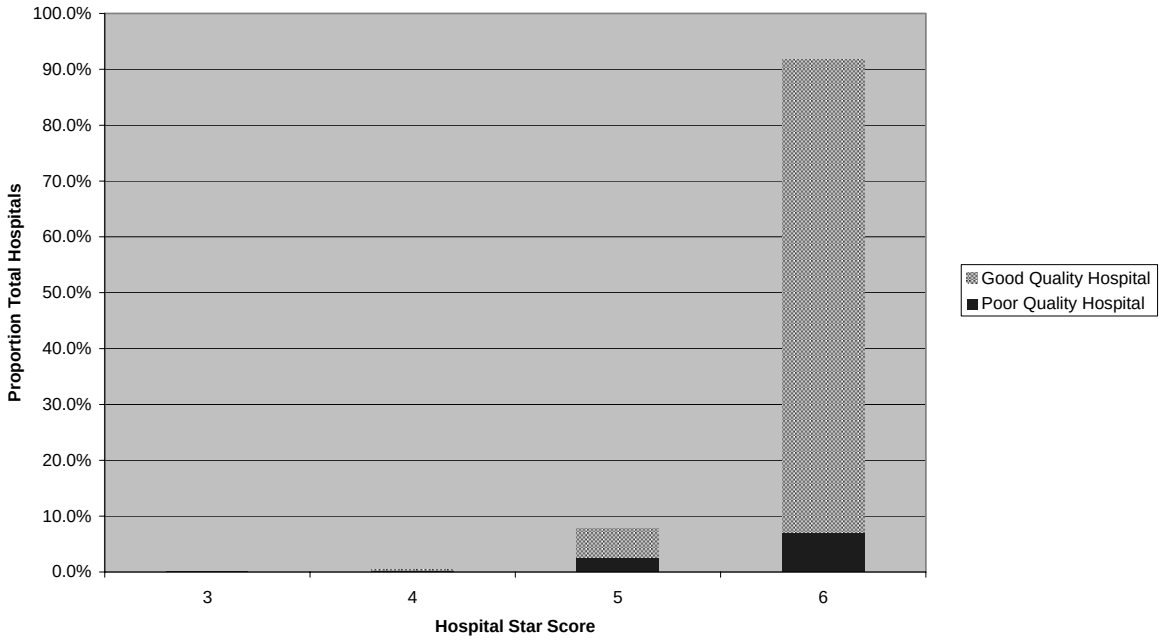


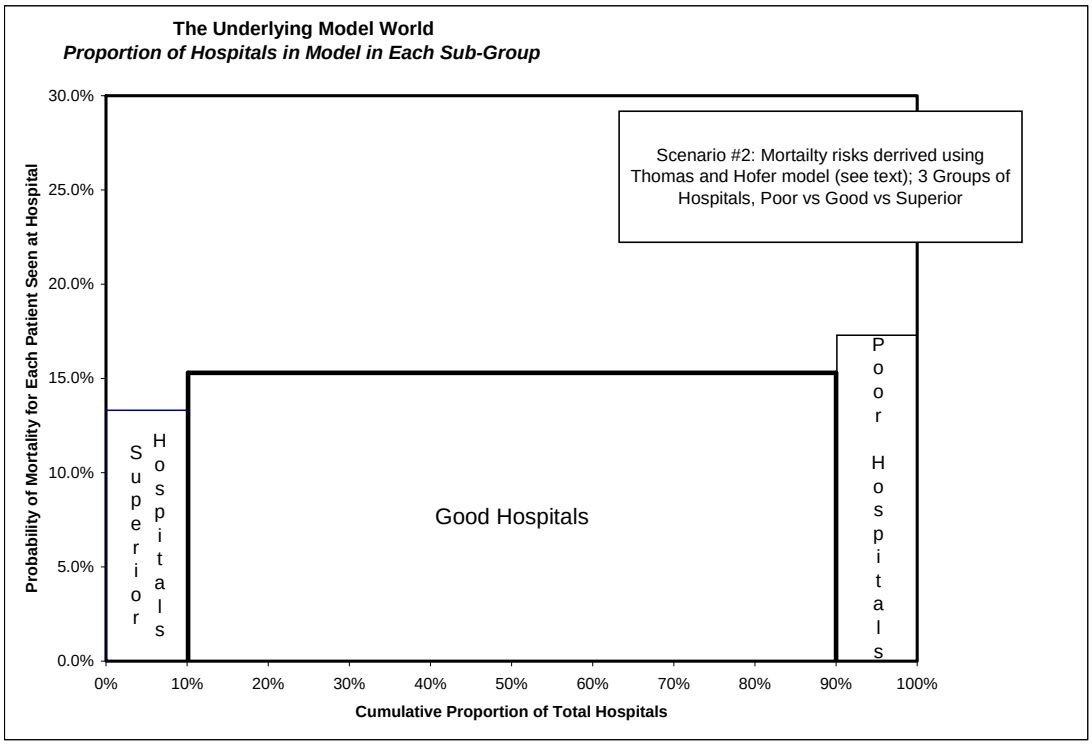
Table 18: Scenario 1: Expected score distribution for good vs. poor hospitals over 3 years

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.

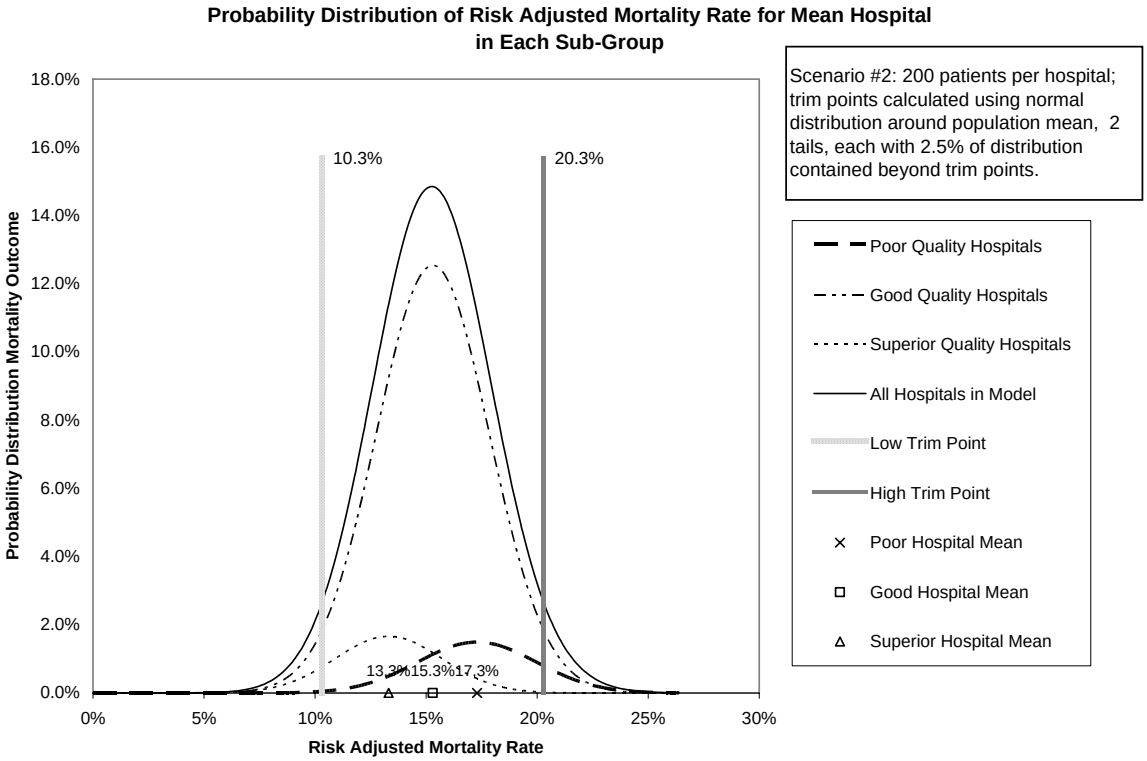
Figure 8: Scenario 2: Hypothetical world of hospitals



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.

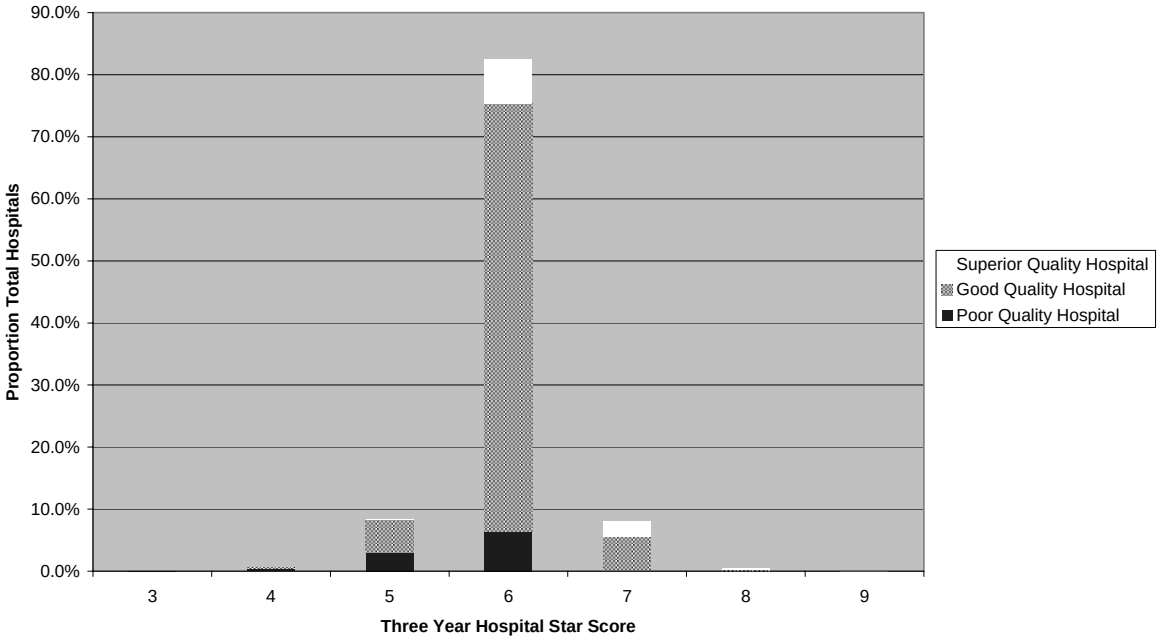
Figure 9: Scenario 2: Hypothetical world and evaluation function



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

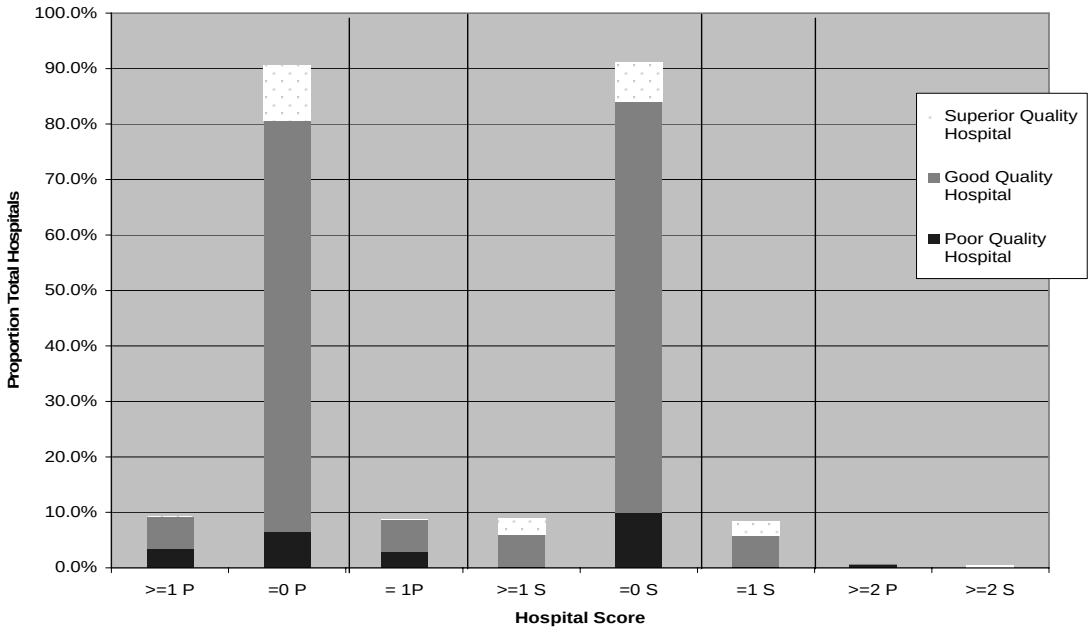


Derivative scores were used to assess whether further discrimination could be obtained among the three sub-groups. The measures are *never poor* (= 0 P), *ever poor* (>= 1 P), *exactly 1 poor* (= 1 P), *mostly poor* (>= 2 P), *never superior* (= 0 S), *ever superior* (>= 1 S), *exactly 1 superior* (=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

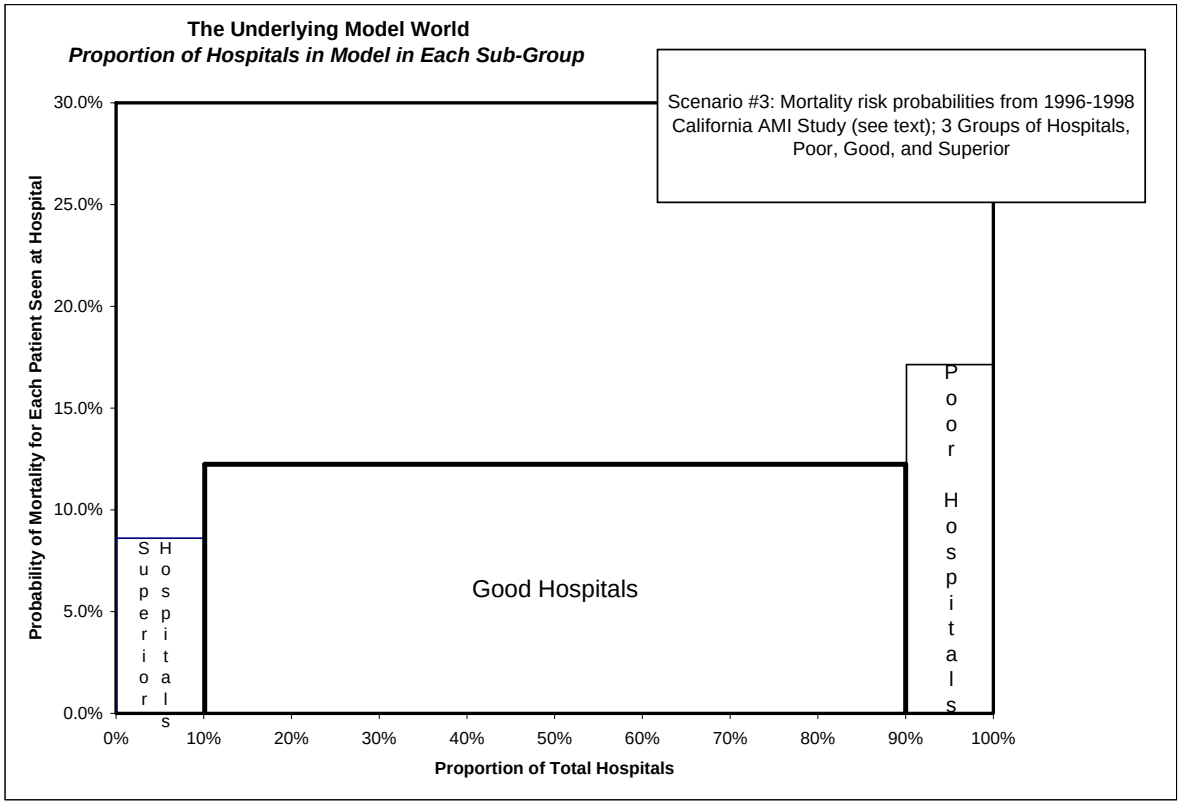
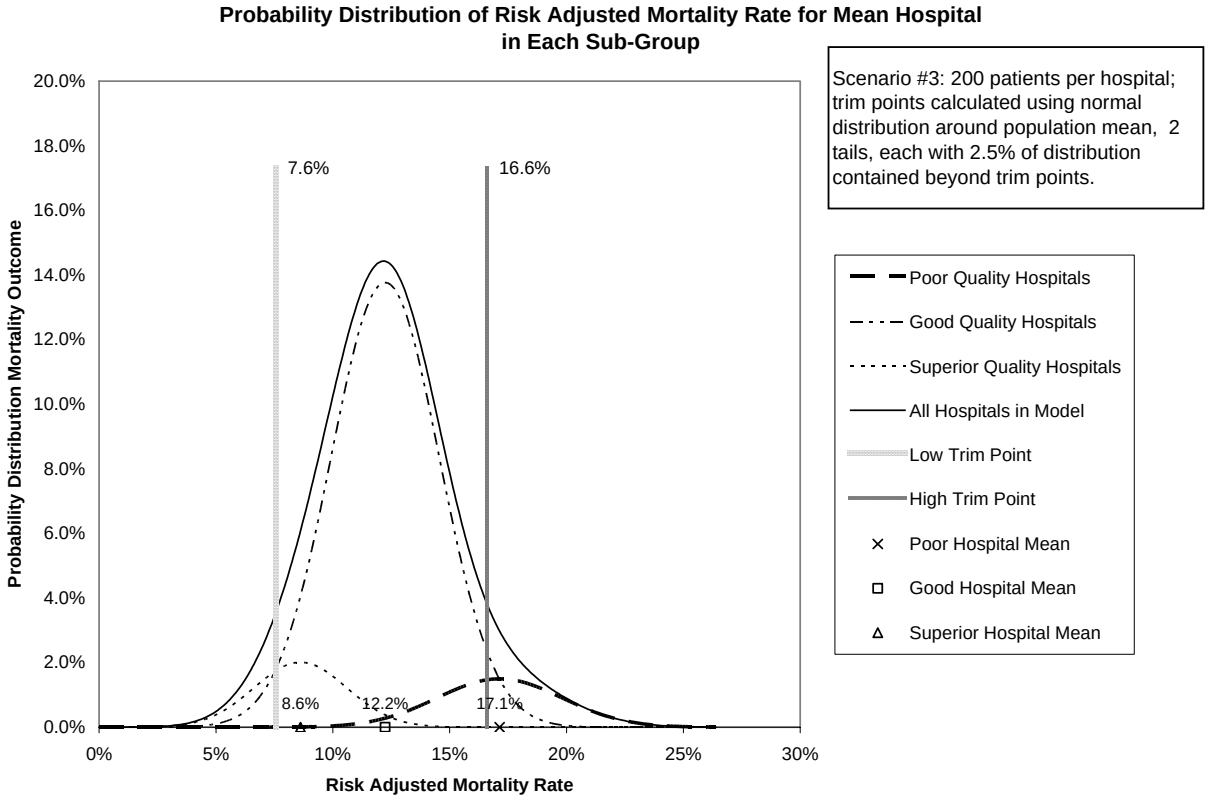
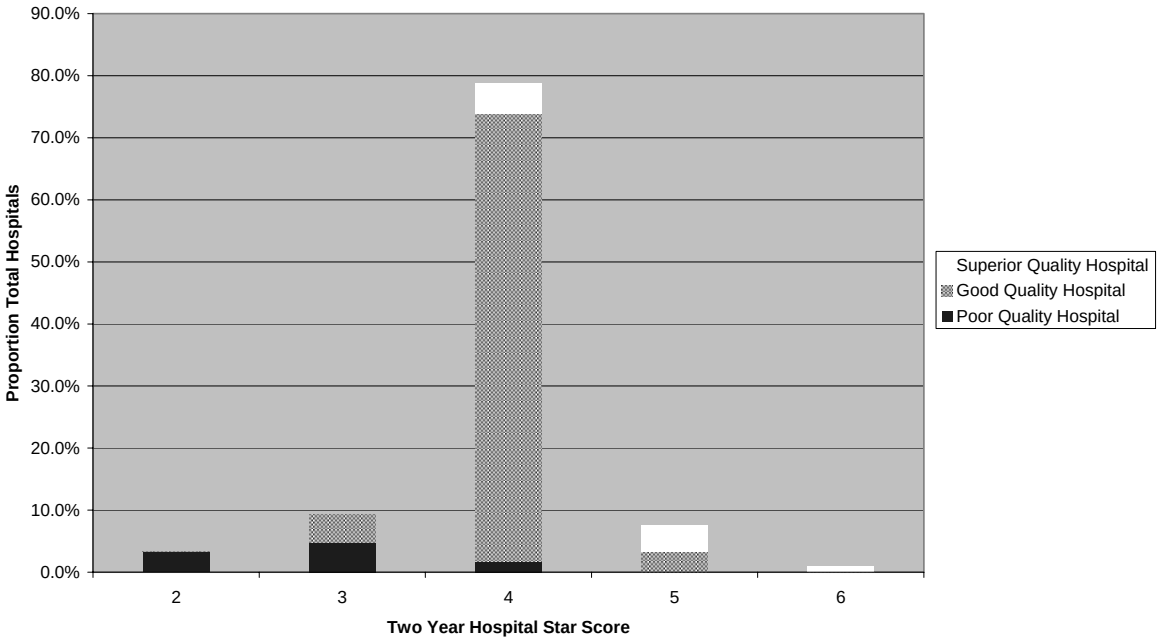


Figure 13: Scenario 3: Hypothetical world and evaluation function



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

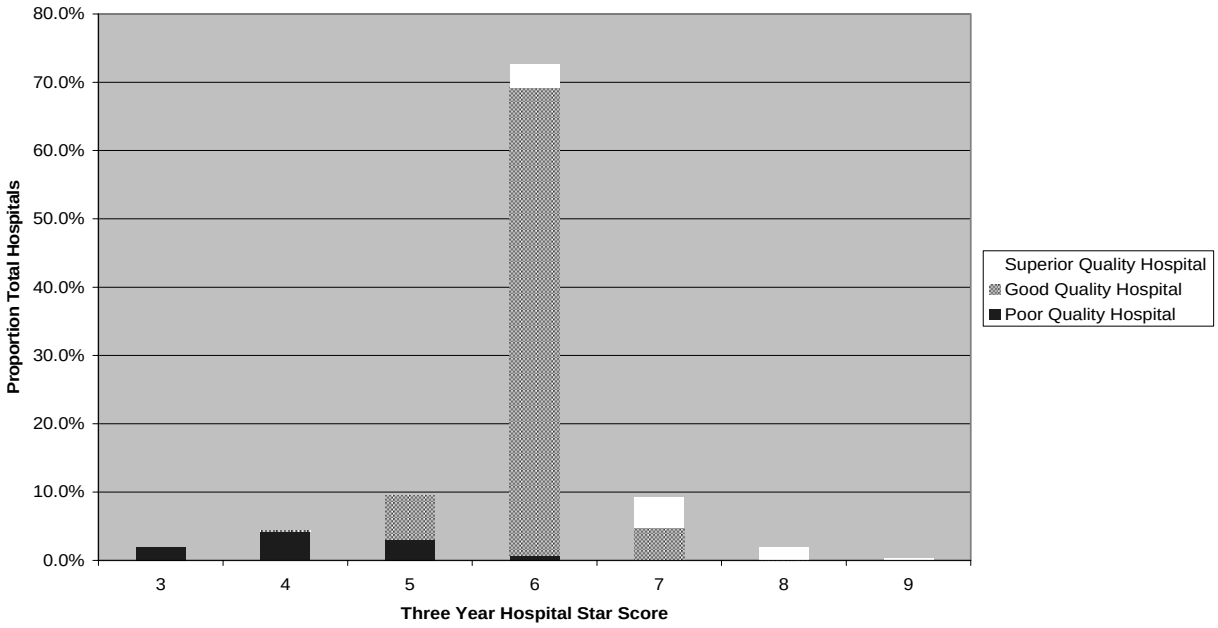


Figure 16: Scenario 3: Three-year derivative scores, predictive values

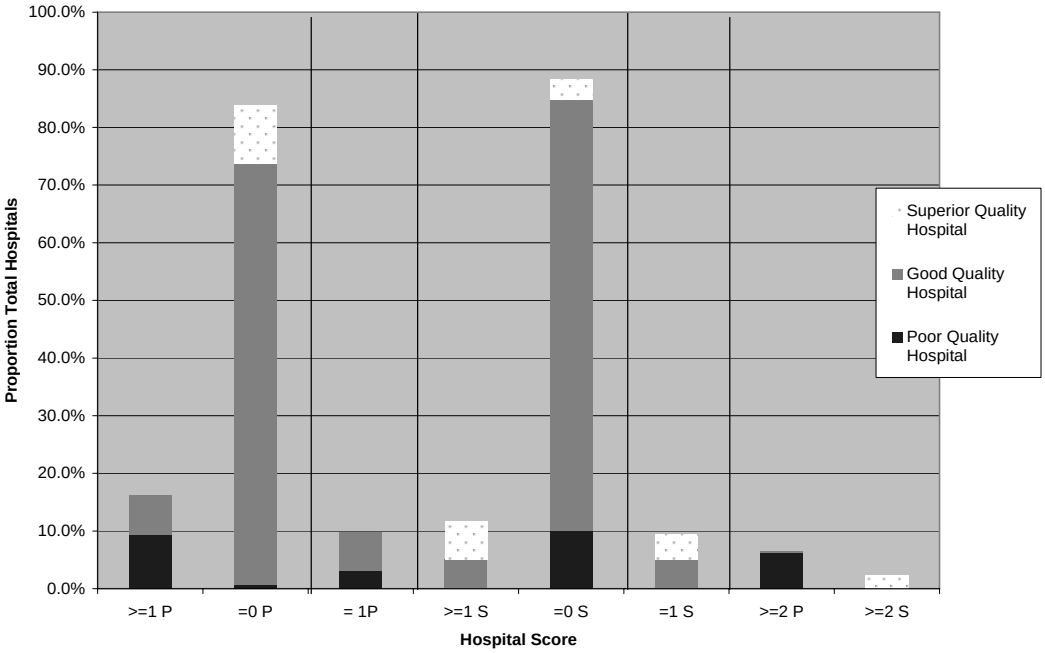
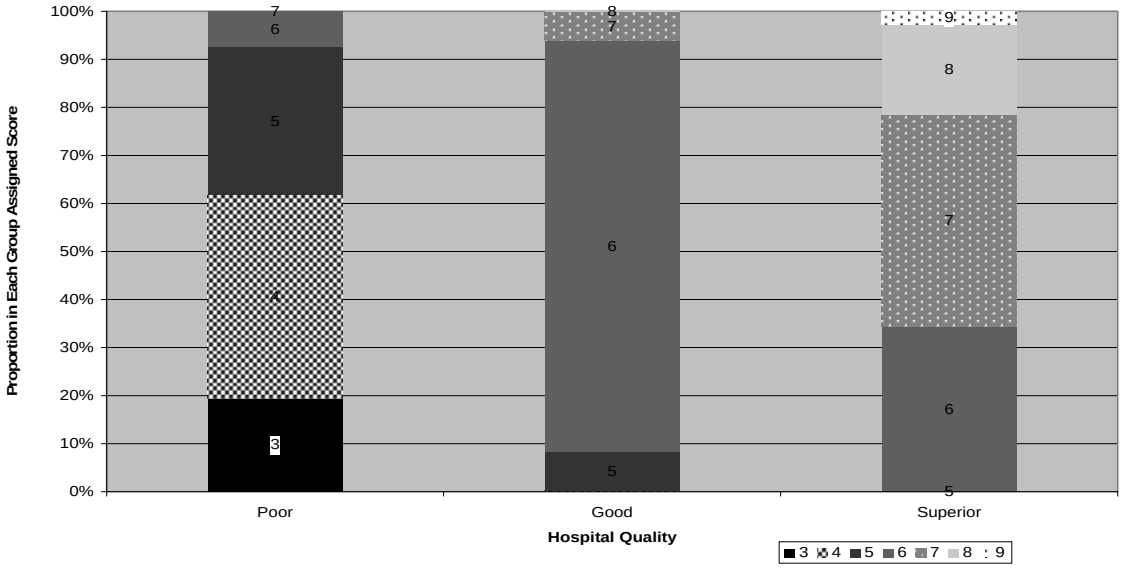


Figure 17: Scenario 3: Distribution of 3-year derivative scores, predictive values



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

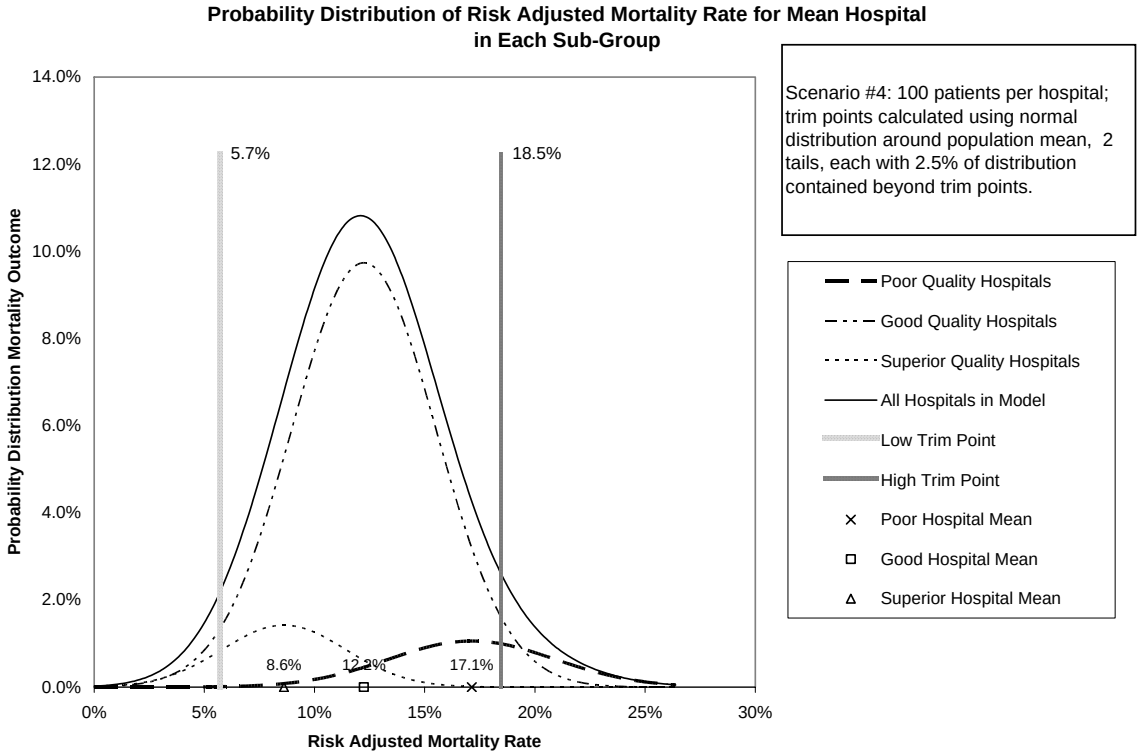
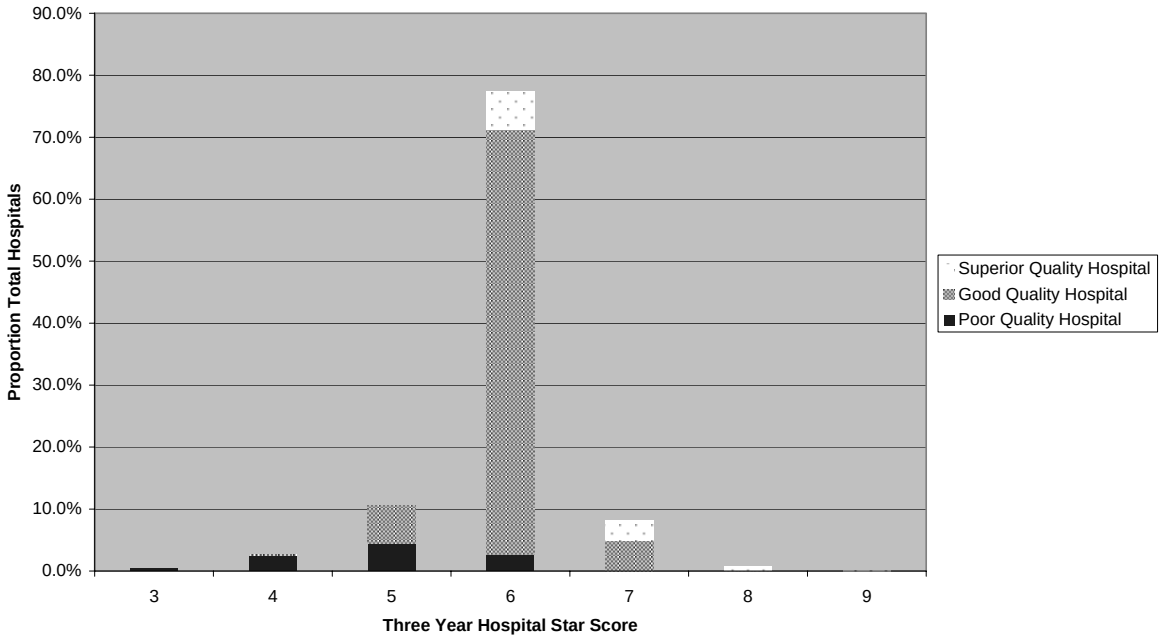


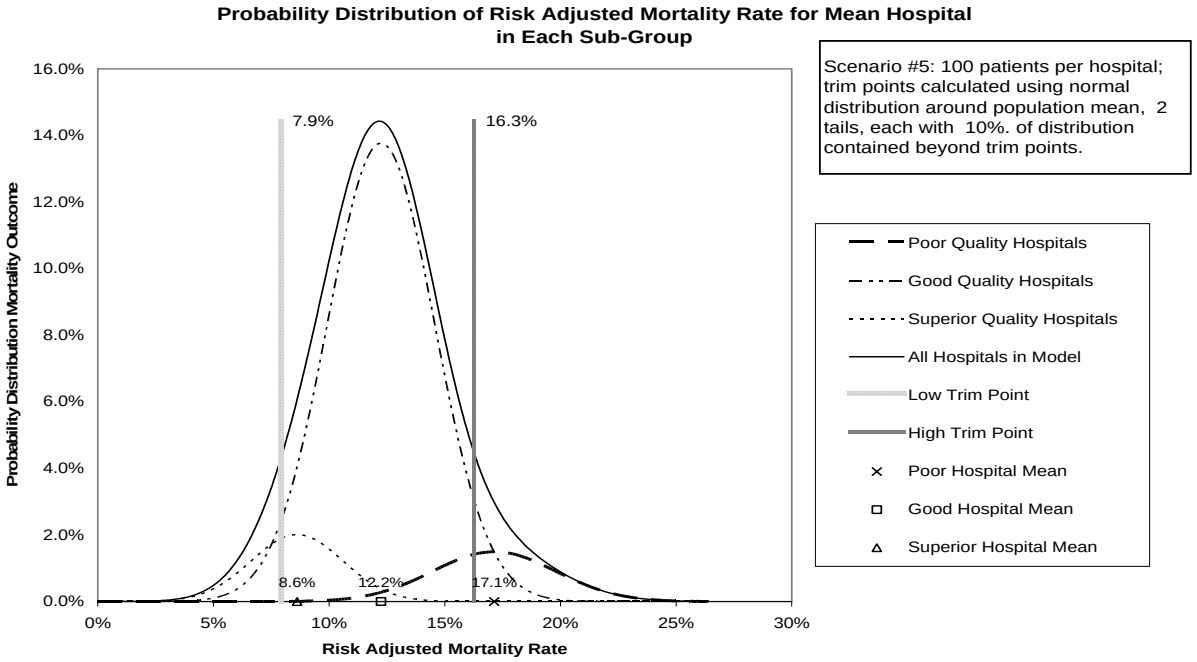
Figure 19: Scenario 4, year 3: Proportion of superior, good, and poor hospitals by 3-year star score



Scenario 5: Identifying a Higher Proportion of Outliers

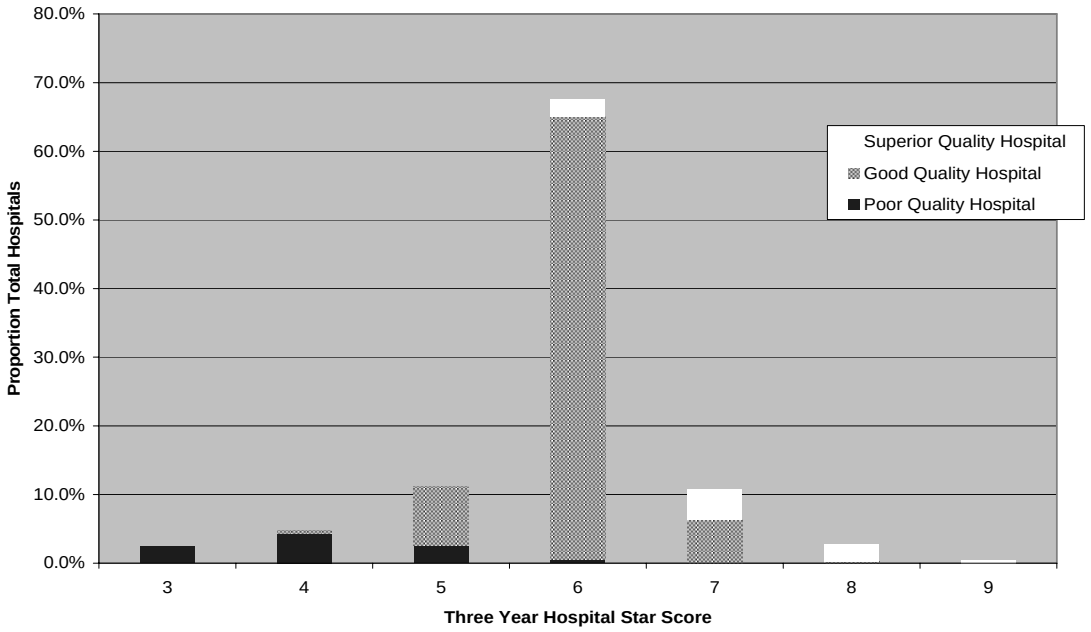
In this simulation, the same hypothetical world as in scenario 3 was used, however, the definition of the trim points for the grading function was changed. In this scenario, the trim points are set such that 10% of the overall hospital quality distribution lies to the right of the upper trim point, and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