Technical Review
Number 10

Strategies To Support Quality-based Purchasing: A Review of the Evidence

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-02-0017

Prepared by:
Stanford–University of California San Francisco Evidence-based Practice Center

Principal Investigator
R. Adams Dudley, M.D., M.B.A

Investigators
Anne Frolich, M.D.
David L. Robinowitz, M.D.
Jason A. Talavera, B.S.
Peter Broadhead, B.A.
Harold S. Luft, Ph.D.

Other Contributor
Kathryn McDonald, M.M.
EPC Associate Director

AHRQ Publication No. 04-0057
July 2004

Suggested Citation:
Dudley RA, Frolich A, Robinowitz DL, Talavera JA, Broadhead P, Luft HS. Strategies To Support Quality-based Purchasing: A Review of the Evidence. Technical Review 10. (Prepared by the Stanford-University of California San Francisco Evidence-based Practice Center under Contract No. 290-02-0017). AHRQ Publication No. 04-0057. Rockville, MD: Agency for Healthcare Research and Quality. July 2004.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome written comments on this technical review. They may be sent to: Acting Director, Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Acting Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services of a particular drug, device, test, treatment, or other clinical service.

Acknowledgments

We are grateful to our expert advisors and peer reviewers for their thoughtful comments and guidance. During the preparation of this report, Peter Broadhead was supported by The Commonwealth Fund. The views presented here are those of the authors and not necessarily those of The Commonwealth Fund, its directors, officers, or staff.

Structured Abstract

Context: Although evidence of quality problems has been available for years, purchaser interest in quality-based purchasing (QBP) is a recent phenomenon. Furthermore, employers who support quality-based purchasing have expressed uncertainty about how to measure quality, especially outcomes, and what incentives to offer to stimulate performance improvement.

Objectives: The objectives of this project were to develop a conceptual model of how incentives influence provider behavior, to summarize what is known from randomized controlled trials about the effectiveness of different QBP strategies, to describe ongoing QBP research, and to perform simulations to determine whether outcomes reports are too influenced by chance events to be used in QBP.

Data Sources: We used online databases (e.g., MEDLINE®) and bibliographies of retrieved articles for the literature search and government and foundation listings to identify ongoing research. For the simulations, we used data from public reports of myocardial infarction outcomes in California.

Study Selection: For the literature review, we sought studies in which providers had been randomized to an incentive group or a control group. We included only projects involving interventions purchasers could plausibly adopt (payment strategies or public reporting of performance). Studies of interventions that were beyond purchaser purview (e.g., implementing clinical guidelines) were excluded.

Data Extraction: We extracted information about the type of incentive used and the clinical and economic context in which it was applied.

Data Synthesis: We evaluated 5,045 publications. Nine were randomized controlled trials, and many of these did not report key characteristics of the incentive or the context in which incentives were applied. Incentives used included additional fee-for-service, quality bonuses, and public release of performance data. The results were mixed: among the 11 performance indicators evaluated, 7 showed a statistically significant response to QBP strategies while 4 did not. We also found 18 ongoing research projects, none randomized. These will yield data about the approaches to QBP currently in use, provider awareness of and concerns about QBP, and some preliminary estimates of the potential impact of QBP.

Regarding assessments of outcomes reports, we found that, under reasonable assumptions and applications, outcomes reports generate meaningful information about provider performance. Providers with good (expected) performance are unlikely to be labeled as *poor quality* in any given period, and very unlikely to be mislabeled more than once in a 3-year period, even if one allowed approximately 10% of hospitals to be labeled *poor* performers annually. In addition, hospitals with *superior* performance were quite likely to be identified as such at least once in 3 years.

Conclusions: Little is known about the impact of QBP on clinical performance. However, it does appear that basing incentives on measurements of outcomes is feasible without undue risk to the reputation or financial status of good hospitals. Ongoing research will only address some of the gaps in our knowledge about QBP, suggesting that much more additional research is

needed. This should include comparisons of alternative QBP approaches and qualitative assessment of the barriers to and facilitators of quality improvement in response to QBP incentives.

Contents

Tables

Figures

Appendixes for this report are provided electronically at www.ahrq.govv/clinic/epcindex.htm. Scroll through the topic list to select this report.

TECHNICAL REVIEW SUMMARY

Strategies To Support Quality-based Purchasing: A Review of the Evidence

Agency for Healthcare Research and Quality

www.ahrq.gov

The mission of AHRQ is to improve the quality, safety, efficiency, and effectiveness of health care by:

- Using evidence to improve health care.
- Improving health care outcomes through research.
- Transforming research into practice.

Introduction

Deficiencies in quality have been widely documented in the U.S. health care system. A recent component of purchaser response to these data has been the pursuit of quality-based purchasing (QBP). However, purchasers have been uncertain both how to measure quality and what incentives to offer to stimulate performance improvement. Furthermore, there has been dispute in the literature about the validity of quality measures, especially outcomes indicators, and the potential for chance variation in outcomes to unduly influence reported performance. Therefore, despite the release of public reports of providers' outcomes by several States, purchasers have been slow to use outcomes reports to drive QBP policies. Without more information about how to proceed with QBP, purchasers risk investing time, resources, and good will without a reasonable expectation of achieving a good return.

In this report,[1] we sought to describe and evaluate the evidence regarding the effectiveness and potential of QBP strategies to improve the quality of care provided in the U.S. health care system. For this report, QBP is defined as payment or reputational strategies aimed at providers that individual employers, employer coalitions, or government programs could plausibly adopt to stimulate the improvement of quality in health care. With respect to providers, the primary issue within the purchaser's purview is the establishment of incentives—for individual providers or for provider organizations such as medical groups and hospitals—that either stimulate or inhibit provider behaviors to improve quality (strategies aimed at consumers, such as variable copayments, were not considered). Specifically, this report focuses on the two types of incentives in widespread use—*performance-based payment* and *reputational incentives* arising from the public release of performance data.

U.S. Department of Health and Human Services
Public Health Service

[1] This summary was taken from Technical Review 10, *Strategies To Support Quality-based Purchasing: A Review of the Evidence*. The suggested citation for this summary is: Dudley RA, Frolich A, Robinowitz, DL, Talavera JA, Broadhead P, Luft, HS. Strategies To Support Quality-based Purchasing: A Review of the Evidence, Summary, Technical Review 10. (Prepared by Stanford-University of California San Francisco Evidence-based Practice Center under Contract No. 290-02-0017). AHRQ Pub. No. 04-P024. Rockville, MD: Agency for Healthcare Research and Quality. 2004.

Objectives

Because quality-based purchasing is in its infancy, the first objective was to develop a conceptual model of how QBP strategies could be used to create incentives for providers to improve care. The second objective was to identify all the published, peer-reviewed randomized controlled trials of QBP and to summarize what is known about the relative effectiveness of different strategies.

Because the literature on QBP is sparse, a third objective was to identify ongoing research that might increase our knowledge. Finally, since one of the main issues purchasers face is whether to use reports of outcomes of care, the fourth objective was to determine whether outcomes reports convey meaningful information or are too influenced by chance events to be useful.

Conceptual Model

There is extensive theoretical literature about the determinants of the effectiveness of incentive arrangements in several disciplines, including economics, psychology, and organizational behavior. An expansive review of that literature is beyond the scope of this report. However, this research has pointed out, among other things, the influence of the characteristics of the incentive itself and of the context in which it is applied on the likelihood that the incentive will be effective.

- **Characteristics of the incentive.** Important financial characteristics include whether it is directed to the optimal recipient. Recipients could include, for instance, the individual provider, provider groups, or even community organizations, with "optimal recipient" varying depending on the goal and degree of coordination among providers required. Other important financial factors are the potential impact on revenue (based on the magnitude of the incentive and the proportion of encounters or patients to which it applies) and the cost of complying with the performance measure.

  Nonfinancial characteristics are more numerous and subtle. These include perceived attainability of the performance goals set, the acceptability of those goals (their congruence with professionalism, altruism, and intrinsic motivation and with provider preferences for domain of performance measured), and the approach to reinforcement (e.g., positive vs. negative reinforcement).
- **Contextual factors.** Although these factors are likely very important, they have received little attention, especially in the empirical literature. In particular, we posit that there are predisposing factors—such as the mix of other incentives in the market and individual provider characteristics or a provider organization's understanding of its mission—that that will determine the likelihood of a provider having any interest in responding to a newly introduced QBP program. Furthermore, we also hypothesize that there are enabling factors—especially at the organization level, where many aspects of the structure of care are determined, and at the patient level—that will facilitate or inhibit any efforts a provider makes to improve care.

In emphasizing both the characteristics of the incentive itself (the QBP stimulus to improve) and the predisposing and enabling factors that may vary among providers and markets, we believe this model complements and can integrate most of the existing theories of incentives. It is offered simply to ensure that adequate consideration is given to all key factors in designing both studies of quality-

based purchasing and future QBP programs.

Methods for the Literature Search and Identification of Ongoing Research

Literature Searches

To be considered an article that provided evidence regarding QBP, the intervention in the trial had to be a performance-based payment or reputational incentive strategy that could plausibly be introduced by a purchaser. The focus was on articles that provided definitive primary data from randomized controlled trials, because most non-randomized designs in this domain are severely confounded, especially by selection bias in which providers were willing to accept new incentives, regression to the mean (since organizations may have chosen to introduce incentives targeted at problem areas that would have improved anyway), the Hawthorne effect, and other sources of variation in performance over time not related to the incentive. Articles that did not have clear inclusion and exclusion criteria and greater than 75% followup were excluded.

Standard search strategies were used. These strategies involved the querying of two online databases (MEDLINE® and Cochrane) using key words, followed by evaluation of the bibliographies of relevant articles, Web sites of relevant organizations (especially of funding agencies providing project summaries and of employer organizations pursuing QBP), and reference lists provided by the Technical Expert Panel. At least two investigators screened titles, abstracts, and articles, as necessary, to determine if they met inclusion criteria. From each included article, the following data were extracted, when available: information describing financial and nonfinancial characteristics of the incentive, financial characteristics of the environment including dominant proportion of income from fee-for-service or capitation and other incentives faced, provider characteristics, organizational capabilities, and patient factors, as well as references in the bibliography that might meet inclusion criteria.

Identifying Ongoing Research

The online databases HSRProj and GOLD—the Grants-On-Line Database of the Agency for Healthcare Research and Quality (AHRQ)—were searched, as well as the Web sites of other funders or coordinators of projects (e.g., the Leapfrog Group). Finally, staff at AHRQ, the Robert Wood Johnson Foundation (RWJF), the California HealthCare Foundation (CHCF), and the Commonwealth Fund were asked whether ongoing research that met the inclusion criteria was being funded by those organizations. Two investigators reviewed the abstracts of projects identified from the database searches to assess relevance to the Technical Review. Discrepancies in inclusion were resolved by discussion and re-review and by discussion with project officers at funding agencies or with the principal investigator of the project under consideration.

Results From the Literature Search and Identification of Ongoing Research

Articles Included in the Literature Search

The literature searches identified 5,045 unique candidate articles for inclusion, of which 4,882 were eliminated after review of their abstracts. The remaining 163 articles underwent full text review. Among these there were only nine randomized controlled trials, eight using performance-based payment as the intervention and one using reputational incentives.[1-10]

Completeness of the Literature

In every article reporting the results of a randomized controlled trial of performance-based payment incentives, there were significant variables from our conceptual model that were either not reported at all or that were incompletely described. The only variables that were reported in all trials were characteristics of the incentive itself: the recipient of the incentive, its magnitude, and the domain of performance measured. Several potentially critical variables were never reported in any trial, including payment incentive as a proportion of total income, the costs of complying with the incentive, and most enabling factors at the organizational level.

Findings From Trials of Performance-based Payment

The eight trials of performance-based payment were neither consistent in their design of the independent variable (the financial incentive offered) nor comparable in terms of their dependent variable (the performance indicator measured). Thus, their results are presented as a function of several of the variables within the conceptual model (those that are actually reported for all papers). In total, ten hypotheses and ten dependent variables were tested because one study had two intervention arms (a fee-for-service arm and a bonus arm) compared to controls, and one had two dependent variables (screening for smoking and smoking cessation).

Recipient of incentive. In four studies, the recipient of the incentive was an individual provider, while in the other four the recipient was the provider group or could be either an individual provider or a group. Among the studies targeting individual providers, there were five positive and two negative results; among the studies in which the target was or could be the provider group, there were one positive and two negative results. (In general,

the term "positive" is used to mean an effect in the desired direction—the incentive worked—and "negative" to mean there was no significant effect of the incentive on the outcome measure.)

In seven studies, with a total of nine dependent variables, the target of the incentive was a physician. Of the nine dependent variables assessed, five showed a significant relationship to the incentive in the expected direction and four showed no significant change after the incentive was introduced. A single study involved pharmacists and was positive.

Magnitude of the incentive. Incentives ranged in magnitude from $0.80/flu shot to a bonus of up to $10,000 per clinic per year. There was no consistent relationship between the magnitude of the incentive and response (though the lack of similar interventions and dependent variables make it unlikely that any pattern could be detected, even qualitatively).

Fee-for-service vs. bonus. There were five dependent variables in fee-for-service studies (that is, the intervention involved paying providers a higher than usual fee for each encounter if and only if a performance standard was met) and five in bonus studies. Among the fee-for-service studies, four were positive and one was negative. Among the bonus studies, two were positive and three were negative.

Performance domain measured. Among the articles included, there were seven studies of preventive care with nine dependent variables assessed. Among these nine outcomes, five were positive and four were negative. The single study addressing chronic care was positive.

Patient factors. Authors did not report the burden adherence would place on patients in any of the articles. However, in a general sense, incentives to achieve performance were found to be more effective when the indicator to be followed required less patient cooperation (e.g., receiving vaccinations or answering questions about smoking) than when significant patient cooperation was needed (e.g., to quit smoking).

Findings From Trials of Reputational Incentives

There was only one randomized controlled trial of reputational incentives. This study showed that hospitals with low performance scores were more likely to engage in quality improvement activities. This was especially true for hospitals whose performance was released to the public (as opposed to being kept confidential).

Ongoing Research Identified

We identified no currently ongoing randomized controlled trials of QBP strategies from any funding source. There were 18 ongoing research projects about QBP. For many of these, the exact nature of the performance measures and the incentive were still being determined. For some, the study design is observational; that is, health plans are making decisions about incentives without input from the investigators, but the investigators are assessing the response.

Expected Knowledge To Be Gained From Ongoing Research

Ongoing research being conducted by AHRQ, the Robert Wood Johnson Foundation, the California HealthCare Foundation, and the Commonwealth Fund will provide some important additional information about quality-based purchasing. For example, several studies will describe the type and frequency of use of QBP strategies; others will investigate provider reactions to incentives in terms of willingness to participate in programs and awareness of the incentives offered. In addition, some investigators will obtain quantitative and qualitative information about attitudes towards incentives used and performance targets set (such as salience, clinical validity, and whether the performance measures were within the providers' scope of control). These studies may be useful for understanding providers' motivation to respond and organizational decisionmaking when incentives are offered. Still other projects will report on the tools used to communicate incentives, rather than the provider or consumer response to the incentive.

The Rewarding Results projects (with components sponsored by RWJF, CHCF, and AHRQ) as well as several others will provide assessments of the impact of incentives on traditional performance measures of structure, process, and outcomes. Although none of these is randomized and all involve organizations that self-select to adopt or participate in incentive programs, taken together they will provide preliminary evaluations of QBP in Medicaid, Medicare, and commercial insurance settings and will cover many different approaches to incentives.

Among the interventional studies, there are also some major differences in the characteristics of the incentives themselves between the prior literature and the ongoing research. For instance, the ongoing studies involve actual health plans or government programs making an ongoing commitment to an incentive strategy, rather than a researcher making a short-term payment intervention (which was the situation in the prior studies). Similarly, all the studies included in the literature review above involved incentives directed at only a small number (usually just one) performance indicator for a single condition or type of patient. However, all the ongoing interventional studies identified involve multiple measures (often ten or more) across a variety of conditions and distinct patient populations. Both these factors—that the incentive comes from a payer (e.g., health plan, government)

and that there are multiple quality indicators—will provide more broadly applicable evidence about the probability that provider investments in quality improvement (e.g., installing a new information system) can be recouped relative to previously studied incentive strategies.

Methods for Simulations To Assess the Usefulness of Outcomes Reports

To examine the role of random variation versus true hospital quality differences in assessing reported hospital outcomes, simulations were developed to determine how often hospitals would be mislabeled in public reports. To do this, first assumptions were made 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. 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. (This is discussed in terms of mortality rates, but the same logic applies to any other outcome.) 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).

Assumptions for the Simulations

Prior studies have suggested that the influence of chance is very great, perhaps enough to cause outcomes reporting to do more harm than good. However, these were based on assumptions—usually based on implicit reviews of overall performance rather than explicitly assessing compliance rates for specific aspects of care—that included a relatively simple performance distribution (e.g., only "good" and "bad" hospitals) with small differences in performance between the groups.[11] For completeness sake, some simulations were performed using assumptions taken from prior research. However, some simulations were done in which assumptions about hospital performance were based on published California data about acute myocardial infarction mortality rates from 1991-1998. These data showed approximately 10% of hospitals had been labeled "better than expected," 80% had been labeled "no different than expected," and 10% had been labeled "worse than expected" in most years. 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. Estimates were obtained of probability of death at poor, good, and superior quality hospitals using 3-year grouped data from the published California study of acute myocardial infarction outcomes. Hospitals that were found consistently—i.e., over two or three of the 3-year periods included in the data (1991-1993, 1994-1996, and 1996-1998)—to have statistically significantly higher than expected mortality were included in the group of poor hospitals; those with consistently lower than expected mortality were included in the

group of superior hospitals, and all others were in the good or expected group.

Assessments of Outcomes Reports and Labels

Using these assumptions, simulations were run to determine the proportion of hospitals from each group (i.e., hospitals that were truly poor, good, or superior) that would be designated into each group (i.e., the proportion that would receive the labels "poor," "good," or "superior"). Since hospitals that have generally been performing well which have a single event in which they are labeled "poor" might face few consequences, simulations were performed not just for a single point in time, but also for two or three measurement periods. The impact of varying sample sizes at a hospital was also considered.

Results From Simulations To Assess the Usefulness of Outcomes Reports

Simulations Using Assumptions From the Literature

As expected, when the assumptions used previously are made again, the results suggest that random variation causes frequent mislabeling of hospitals in a single period, with potentially more than half the hospitals labeled "poor" actually coming from the population of good hospitals. However, when the analysis is extended over as few as 3 years, mislabeling more than once becomes extremely unusual for good hospitals; fewer than 0.2% of good hospitals would have this outcome even if one assumes small mortality differences between poor and good hospitals.

Simulations Using Assumptions From California Data

The mortality rates for acute myocardial infarction for poor, good, and superior hospitals in California in 1996-1998 were 17.1%, 12.2%, and 8.6%, respectively. Using these mortality rates, superior hospitals were almost never labeled "poor" and vice versa. Over a 3-year period (with reports each year), 92.5% of poor hospitals would be labeled as such at least once (vs. only 8.7% of good hospitals) and almost all the hospitals that were labeled poor more than once would in fact be poor. Similarly, most superior hospitals would receive at least one such label, and almost all hospitals labeled superior more than once would actually be superior.

Discussion and Future Research

Quality-based purchasing is a relatively new topic, and very few studies were found that address the key questions about QBP. Comparison of our conceptual model to the available research also points out that the studies available are incomplete in their reporting of potentially key mediators of the effects of incentives.

Nonetheless, there is evidence that, in some circumstances, both performance-based payment and reputational incentives can work. Preliminary evidence suggests that, consistent with theory, the revenue potential from incentives and the costs of achieving performance goals may influence response, as will enabling or inhibiting factors at the patient level. In addition, ongoing research will inform us about the extent of use of QBP, provider attitudes toward both incentives and the use of various types of performance measures, and preliminary estimates (though the data will come from non-randomized studies) of the impact of QBP on quality.

Much additional research is needed, including both qualitative and quantitative designs. Since randomized trials are expensive and providers often will not agree to randomization, funders might consider looking for natural experiments or situations in which non-randomly selected control groups could reasonably be used (as when a health plan decides to roll out a QBP approach first in one city, then in another; of course, even in these situations there will probably be a reason as to why one city was chosen to be first that could bias results). One such example may be the recently initiated Centers for Medicare & Medicaid Services Premier Hospital Quality Incentive Demonstration that recognizes and provides financial rewards to hospitals that demonstrate high quality performance in a number of areas of acute care (see: www.cms.hhs.gov/researchers/demos/phqidemo.asp).

Furthermore, subsequent research should explicitly address the elements from conceptual models that have largely been ignored. Investigators should address the reality that while much of performance is ultimately determined by the actions of individual providers, enabling factors at the organizational and community levels that determine the structure and processes of care are also important and could be targets for incentive strategies. In addition, studies that address the combination of performance-based payment with reputational incentives are needed.

Finally, one must recognize that a prominent barrier to QBP is that the science of performance measurement is still underdeveloped. Purchasers interested in QBP have limited choices for performance measures and these disproportionately target preventive care and structure or processes rather than outcomes. That is, the available set of metrics is not broadly representative of all care, while purchasers must pay for care across the entire clinical spectrum. This suggests that research into QBP should be accompanied by further development of the basic tools of performance measurement.

Conclusion

The environment in which purchasers and providers interact is rapidly changing. There is clearly growing interest in QBP and some evidence that both payment and reputational incentives can work; but, to date, there is little unequivocal data on which to base QBP strategy selection. Our modeling suggests that, with appropriate caution, outcomes measures can be included among the performance indicators used for QBP. Furthermore, the notion of using incentives to encourage high quality (as well as actually measuring quality) is much more acceptable than it was a few years ago, and this has increased the number of opportunities to study QBP. Researchers have responded with a broad portfolio of ongoing research that promises to both outline current trends in the use of QBP and offer some preliminary evaluations of several different incentive approaches. Additional policy-relevant research, including studies incorporating in their designs conceptual considerations such as those outlined here, may rapidly advance our understanding of how to use performance measurement and incentives to improve the quality of health care Americans receive.

For More Information

Printed copies of the Technical Review from which this summary was taken may be obtained free of charge from the AHRQ Publications Clearinghouse by calling 800-358-9295. Requesters should ask for Technical Review 10, *Strategies To Support Quality-based Purchasing: A Review of the Evidence* (AHRQ Pub. No. 04-0057).

Additionally, the Technical Review and this summary will be available online through AHRQ's Web site at www.ahrq.gov.

References

1. Christensen DB, Holmes G, Fassett WE, et al. Influence of a financial incentive on cognitive services: CARE project design/implementation. *J Am Pharm Assoc*. Sep-Oct 1999;39(5):629-639.
2. Christensen DB, Hansen RW. Characteristics of pharmacies and pharmacists associated with the provision of cognitive services in the community setting. *J Am Pharm Assoc*. Sep-Oct 1999;39(5):640-649.
3. Davidson SM, Manheim LM, Werner SM, Hohlen MM, Yudkowsky BK, Fleming GV. Prepayment with office-based physicians in publicly funded programs: results from the Children's Medicaid Program. *Pediatrics*. Apr 1992;89(4 Pt 2):761-767.
4. Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. *Ambul Pediatr*. 2001;1(4):206-212.
5. Hickson GB, Altemeier WA, Perrin JM. Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study. *Pediatrics*. Sep 1987;80(3):344-350.
6. Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. *Am J Public Health*. Nov 1998;88(11):1699-1701.
7. Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. *Pediatrics*. Oct 1999;104(4 Pt 1):931-935.

8. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. Performance-based physician reimbursement and influenza immunization rates in the elderly. The Primary-Care Physicians of Monroe County. *Am J Prev Med*. Feb 1998;14(2):89-95.
9. Roski J, Jeddeloh R, An L, et al. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. *Prev Med*. Mar 2003;36(3):291-299.
10. Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff* (Millwood). Mar-Apr 2003;22(2):84-94.
11. 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.

AHRQ Pub. No. 04-P024
July 2004

Technical Review

1. Introduction

Background

Deficiencies in patient safety and quality are rife in the U.S. health care system.[1-3] Although evidence of quality problems has been available for many years, purchaser initiatives to ensure that beneficiaries receive high quality care have become common only in the last few years.[4, 5] As they have begun to pursue or consider quality-based purchasing (QBP), some employers have expressed uncertainty about what information to use to measure quality and what incentives to offer to stimulate performance improvement, and have expressed frustration at the difficulty of implementing QBP.[5] Furthermore, there has been dispute in the literature about the validity of quality measures, especially outcomes indicators, and the potential for chance variation in outcomes to unduly influence reported performance.[6-8] Therefore, despite the release of public reports of providers' outcomes by several states, purchasers have been slow to use outcomes reports to drive QBP policies.[5, 9, 10]

In fact, purchasers have historically focused more on price than quality when making health care purchasing decisions.[4, 11] Recently, however, both private[12, 13] and government purchasers[14] in the United States have committed to improving quality. In addition, the trend of using incentives to stimulate improvement has spread to other nations as well.[15, 16] In the absence of good information about how to proceed with QBP, however, purchasers risk investing time, resources, and good will without a reasonable expectation of achieving a good return.

Over the last several years, several important studies and reviews have been published that offer some insight into how QBP strategies such as offering financial incentives to providers or the provision of performance data to providers can influence quality of care. Unfortunately, many of these studies have examined only one or a small number of factors that could have an impact on performance and there have been no prior attempts to bring all elements together into a single comprehensive description of how to do QBP.

The nomination of QBP for an evidence report was submitted by the Employer Health Care Alliance Cooperative (The Alliance). Through discussions between Agency for Healthcare Research and Quality (AHRQ) and the Alliance and based on a feasibility report prepared by the EPC, AHRQ determined that a comprehensive review of the QBP literature and ongoing research could provide insights about the current state of the art in QBP. In addition, in light of the uncertainty about the value of measurements of providers' outcomes, the Agency determined that the literature review should be supplemented by explicit consideration of the potential validity of outcomes reports and whether risk adjusted outcomes are too severely influenced by chance events to be valid measures used in QBP.

Purpose of This Report

The purpose of this report is to describe and evaluate the evidence regarding the effectiveness and potential of QBP strategies to improve the quality of care provided in the U.S. health care system. For this report, QBP is defined as purchasing approaches that individual employers, employer coalitions, or government programs could plausibly adopt to stimulate the improvement of quality in health care. The issue of *plausible* purchaser adoption is critical. There are many potential approaches to improving the quality of care, but most are beyond the

control of purchasers. For example, the creation of a set of guidelines for the provision of diabetes care or the establishment of a team to make antibiotic recommendations may be highly valuable approaches to improving quality, but are not purchaser functions and would not be strategies purchasers could implement. Rather, the primary issue within the purchaser's purview is the establishment of incentives—for individual providers or for provider organizations such as medical groups and hospitals—that either stimulate or inhibit provider behaviors to improve quality.[17] (Strategies aimed at consumers such as varying copayments based on provider performance have rarely been studied. In developing key questions with AHRQ and the Technical Expert Panel, it was decided to focus on the purchaser-provider relationship.) Therefore, this report addresses the use of QBP to create provider incentives, the scope of which will be described in the next section.

Because QBP is in its infancy, the first objective was to develop a conceptual model of how QBP strategies could be used to create incentives for providers to improve care. The second objective was to identify all the published, peer-reviewed randomized controlled trials of those incentive systems that purchasers could plausibly adopt and to summarize what is known about the relative effectiveness of different QBP strategies, with a focus when necessary on what the conceptual model suggests is missing from extant literature.

Because the feasibility report for this project had shown that the literature was limited but the questions were timely, a third objective was to identify ongoing research that might increase our knowledge. Finally, since one of the main issues purchasers face is whether to use reports of outcomes of care, the fourth objective was to determine whether outcomes reports convey meaningful information or are too influenced by chance events to be useful.

Rationale for Focus on Randomized Controlled Trials

Our focus was on randomized, controlled trials, because non-randomized designs in this domain can be severely confounded. Potential sources of confounding include selection bias in which providers were willing to accept new incentives, regression to the mean (since organizations may have chosen to introduce incentives targeted at problem areas that would have improved anyway), the Hawthorne effect, and other sources of variation in performance over time not related to the incentive.

To illustrate this point, we consider one of the randomized trials we did include, a study by Hillman et al. performed in Philadelphia in 1993-1995.[18] In this study, the intervention group nearly doubled its rates of cancer screening over the course of the study, but the control group more than doubled its rates, leading to the conclusion that the incentive itself had no effect. The authors conclude that the increase in performance for both groups may have been related primarily to local and national efforts to improve screening rates, rather than to the QBP incentive.

Had this study not been had a randomly selected control, one might have concluded that the incentive worked, and actually had a large effect (since screening increased so dramatically). This could even have occurred if the there had been a non-randomly selected control group, say in Pittsburgh, if the main force causing the increase in screening was local initiatives in Philadelphia to improve care.

In fact, to the extent that one studies natural experiments in which a health plan or government program implements a QBP program in one geographic area but not another or with a particular group of providers but not others, selection bias is almost certain to be present and

potentially significant. This is because purchasers will want to use their resources wisely and will consider, if they cannot implement QBP in all areas, the likelihood of success in one area versus another. They would have an incentive, in fact, to choose areas in which they expect success and to avoid areas in which implementation would be difficult or likely to fail.

Furthermore, it is unlikely that purchasers would be willing to make only the QBP intervention the sole change in a given market throughout the course of the study (most of the ongoing research projects are three or more years long, considering the time for project planning to grant submission through project completion). Judgment would be used to decide which interventions to introduce and where. Thus, if purchasers had introduced a QBP program in an area at one point in time because performance was particularly poor in that region, they might also choose at a subsequent period to invest more in provider education in that area than in a control area in which performance was already better (which may have been what was happening in Philadelphia in the mid-1990s).

As this is an early review of QBP, we considered it very important to avoid misleading potential users. Therefore, after discussions with our Technical Expert Panel and AHRQ staff, we focused on randomized controlled trials only.

Types of Incentives

In the United States, the Institute of Medicine (IOM) has made a compelling case that quality and safety of health care needs to be improved, and recommends that purchasers "align financial incentives with care processes based on best practices and the achievement of better patient outcomes".[17] Furthermore, the IOM also argues that "no payment method is neutral" with regard to quality, in that "efforts to improve quality by correcting overuse, underuse, or misuse all have an impact on provider revenues under all forms of payment".

There are many ways in which payments may influence performance. Much of the focus of research to date has been on the relationship between general approaches to payment, such as fee-for-service (FFS) versus capitation.[19, 20] However, the IOM also proposes basing payment on measurable indices of quality.[17] This approach we refer to as *specific performance-based payment incentives* to improve quality. An example might be a payment of $X for every patient with coronary artery disease whose cholesterol is below some target level (although the performance indicator need not be an outcome, it could also be a structural or process measure).

In addition, the IOM also recommends the communication of provider performance data to the general public and to purchasers. This is also an incentive, either simply because providers care about their reputations or because reputation influences the number of patients a provider organization has or the prices it can charge. Although the public release of performance data clearly could have a financial impact, it could also influence providers in other ways, so we hereafter refer to these strategies as *reputational incentives*.

Incentive Theory

The IOM recommendation about financial incentives draws on principal-agent theory, which addresses relationships in which:

- The two parties have differential abilities and it is therefore desirable for the first party to delegate responsibility for performing a function to the second,

- There is asymmetric information between the two parties, and
- The parties have divergent goals.[21, 22]

These criteria are met in health care, in which patients typically do not have the expertise to determine what care they need or the technical quality of the care they receive. Furthermore, in most instances, care is paid for not by the patient directly, but by a health plan or government health care program. Health plans and government purchasers do not have the clinical expertise or detailed information about each patient to make informed clinical decisions, so they delegate the provision of care to clinicians. In addition, health plans and purchasers cannot measure all the actions providers take that may influence quality of care. Finally, while both health plans and providers care about quality of care, physicians may care more about maximizing income than efficiency, while plans may be more concerned with cost control than quality. In situations such as these, the principal (a health plan or government program) may use incentive payments to encourage its agents (providers) to adopt the principal's goals.

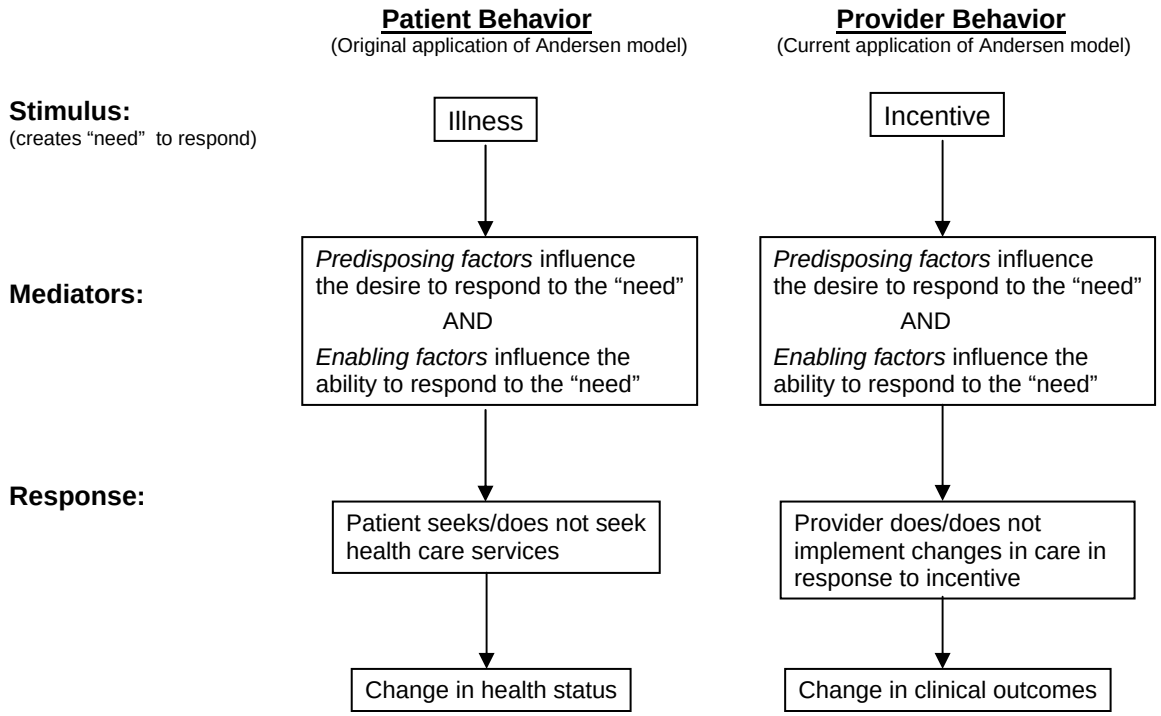
However, other factors besides the relationship between a single principal and its agents may also be critical. The importance of considering the overall financial and nonfinancial milieu in which the agent is acting when designing and implementing financial incentives has been discussed previously,[23, 24] but to our knowledge no conceptual model of the factors influencing the impact of specific incentives on quality has been proposed. Hellinger concludes from a review on the effect of managed care on quality that assessment of any management strategy, which would include incentives, requires detailed information about the characteristics of health plans, providers, and enrollees to draw conclusions.[23] Hutchison et al. point to the importance of considering the context in which financial incentives are designed or implemented to understand their potential effects.[24] The model we propose addresses the reality that the agency relationship between the provider and the health plan or purchaser offering specific incentives occurs in a complex environment in which there are many other potential determinants of provider behavior. Those factors include: the general or predominant way by which the provider is paid, such as FFS, capitation, or salary across all the plans or purchasers with which the provider contracts; the number and character of other incentives; local market factors; organizational characteristics (organizational culture, leadership, etc.); patient characteristics; and physician characteristics.

Since the goal of the provision of QBP incentives is to change provider behavior to improve quality, we believe it is useful to adapt Andersen's Behavioral Model of health care, originally applied to *patients'* behavior in seeking health care services, to *providers'* behavior in deciding to comply (or not) with care according to quality guidelines.[25] The original Andersen model emphasized factors that predispose or enable patients to seek care in response to illness. In economic terms, this is a model of the demand for health care. However, in more general terms, this model offers a fairly flexible approach to placing the behavior of a decisionmaker (the patient) in response to a stimulus (illness) in a broader context (pre-existing factors that predispose or enable a response to the stimulus).

To apply this general approach to providers and QBP, we need only recognize that the provider is a decisionmaker with a stimulus (the incentive the purchaser is offering) who may be more or less predisposed to respond and may encounter have more (or fewer) enabling resources that permit (or inhibit) response. Thus, this application of Andersen's model can be used to address providers' supply of health care and health care improvements (Figure 1). For instance, the demographic characteristics of the individual provider, such as years since the completion of training, may be viewed as predisposing factors toward the provision of specific components of high quality care just as patient demographics have been shown to influence a patient's decision

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

altruism all may vary among providers and could influence one's predisposition to respond to an incentive. Similarly, the ability to monitor behavior and the adoption of reinforcement activities vary among plans. To the extent that providers are aware that they are acting on behalf of a plan that is more able to monitor performance or that has previously engaged in significant reinforcement, they may be more predisposed to respond to the next incentive created. Many of the characteristics of the incentive discussed in either principal-agent or reinforcement theory are also key determinants of the strength of the stimulus to which we show a provider responding and depicted as the "need" to respond in Figure 1.

An important rationale for the use of a conceptual model that integrates a broad array of factors is the possibility to identify variables that have not been adequately studied in the empirical literature. Many of the elements of our model have been identified from a review of health services research literature, but there are aspects of incentives that we believe must be considered but that have received little or no attention. In particular, the essence of an incentive is the net additional income (revenues minus costs) achievable by responding to the incentive. Although the cost to the provider of achieving improved quality is intrinsic to the concept of financial incentives (and thus this point is considered, in the theoretical literature, to be too basic to make), to our knowledge it has not previously been addressed in empirical evaluations of incentives. For that reason, we start with a consideration of the characteristics of the incentive itself.

Characteristics of Incentives

This section describes the potential impact of two key aspects of incentives on provider response: the financial and the nonfinancial characteristics of the incentive.

Financial Aspects of the Incentive

Recipient of the incentive. Incentives can be targeted to individual providers or paid to a provider group or organization (e.g., a medical group or hospital).[18, 32-34] Since changes in clinical process depend on the actions of individual providers, it is conceivable that incentives directed at that level could be more effective than incentives directed to the group. On the other hand, to the extent that improvement requires collective action (e.g., investing in an information system is more feasible if all providers in a group support and participate in the investment; a single provider would find this difficult), incentives may be more effective when directed at the group level.

Revenue potential. Specific incentives can offer a potential increase in revenues (a simple reward) or can involve exposure to risk (e.g., a payment intended to cover all costs associated with an episode, as the Medicare program in the United States creates with its diagnosis-related groups prospective payment). The revenue and profit potential of an incentive are also determined by its structure. For instance, lump sum bonuses for reaching a specified target, bonuses that increase as performance improves (graduated bonuses), or additional FFS payments beyond those usually received (enhanced FFS) are all simple rewards that nonetheless can have very different revenue and incentive implications. In addition, revenue available from the incentive will be affected by whether the performance targets are absolute (e.g., achieve 90% compliance with a guideline) or relative to the performance of other providers (e.g., be among

the top 10% of performers). Finally, it is likely that the salience of the incentive to the provider, and hence the likelihood that it will change provider behavior, is determined at least in part by the proportion of the provider's practice to which the incentive applies. The level to which the incentive is directed—to individual providers or the group or both—may also influence salience.

Impact on cost. Net income from an incentive will also be influenced by the costs to the provider of performing the tasks necessary to improve performance. In general, the total costs will include both the direct costs of doing the activity, complying with the protocol or achieving the outcome, plus the opportunity costs of not doing something else. The relationship between direct cost and improving quality is likely to be complicated, with some fixed and some variable costs, and also to differ depending on the aspects of quality to be improved. There may also be significant start-up, training, or investment costs associated with a change in usual processes, especially if this requires designing new approaches that are not already in use elsewhere. Alternatively, especially if the initial investment required is small, increased quality could also reduce costs.

Responses to incentives, then, will reflect judgments about expected revenues and costs. If the cost of doing X exceeds the return from the incentive, then the incentive will likely fail regardless of its absolute size. It also should be noted that providers' responses will depend on their *perception* of the financial impact of the incentive on their income, not the actual impact. Furthermore, when changes involve up-front costs and downstream benefits, the latter are essentially discounted not just by the usual cost of funds, but also by the perceived likelihood that the bonus payment program will be continued in the future. Few people undertake an exhaustive assessment of the real impact of a changing incentive arrangement, and the actual effect may be obscured by other fluctuations and changes. People tend to respond positively to an incentive if they think it will work for them, and resist it if they do not. So it is quite possible for a QBP program to have a different incentive effect than a rigorous financial analysis would suggest, because the object of the incentive has arrived at a different judgment in his/her own particular way.

Nonfinancial Aspects of the Incentive

Perceived attainability. The extent to which clinicians believe that measured performance is within their control— that is, that they can affect the measure upon which the incentive is based—may be important. Thus, a payment to deliver dietary counseling might result in a higher level of provider response than a payment linked to the number of patients who actually have lost weight at one year, because physicians believe they can influence the former more than the latter.[35] Similarly, requiring a very large improvement relative to prior performance may lead physicians to conclude that the chances of being able to receive the incentive are so small as to not be worth the effort.

Domain of performance measured. The diet and weight loss example highlights the importance of the domain in which performance is measured. Options include:

- Structure—for example, assessing the information technology in place and degree of implementation.[36]
- Processes of care (complying with a defined process)—for example, measuring hemoglobin A1c in patients with diabetes, or the adoption and use of systematic patient recall systems.[18, 32-34, 37, 38]

- Outcomes—for example, achieving intermediate outcomes such as blood pressure control[2] or final outcomes, such as low mortality.

In general, it is easier for providers to control structure or processes than outcomes. This may influence their assessment of their ability to improve measured performance, and hence their willingness to respond to an incentive.[5]

Acceptability of the incentive or performance goal. Grumbach et al. found that physicians were less satisfied with their practice if they faced incentives based on financial outcomes and productivity,[39] which is in accordance with the findings of Hadley et al. and Pantilat et al,[40, 41] and this dissatisfaction with the incentive itself might attenuate response. Physicians with incentives linked to quality of care or patient satisfaction were more likely to be satisfied, perhaps because they found these goals more inherently acceptable than "productivity" for its own sake.

Predisposing Factors

Several factors may predispose providers to respond to an incentive when offered. These include at a minimum the general financial characteristics of the environment (the mix of fee-for-service, salary, and capitation and other incentives used across all payors); traits of the provider; and other characteristics of the market (such as community-wide initiatives to improve performance).

General financial characteristics of environment. There are three main methods of provider payment: fee-for-service, salary (or budget, in the case of an institutional provider such as a hospital or medical group), and capitation.* The dominant financial characteristics of the environment can differ for the organization vs. the individual clinician.[18, 38] For instance, a medical group may primarily receive capitation with occasional FFS payments, but choose to pay each individual provider a salary. Thus, the incentive environment can be different at each level, and hence should be measured and reported for both the group and the individual when possible.

In general the financial incentives inherent in these payment systems are:

- Fee-for-service—financially rewards doing more.
- Salary or budget—payment is independent of activity or outcome, so there are incentives to minimize one's time spent working.
- Capitation—financially rewards doing less of those things that are covered under the capitation payment.

Each of these may modify the effect of a specific incentive, particularly through their influences on opportunity cost.

The potential for opportunity cost is greatest in a *FFS* environment. For example, the opportunity cost of doing more immunizations may be foregoing the performance of activities that generate more fees per unit time. In addition, considerations of opportunity cost may not be confined to simply the relative marginal revenue of an immunization versus a consultation. If immunizing a child is a one-time activity that is unlikely to lead to much subsequent repeat business, while seeing a new elderly patient with a chronic health problem may result in many

* These are archetypes, because in practice, a physician rarely, if ever, receives 100 percent of payments in only one of these forms.

further consultations, the provision of immunizations may have an opportunity cost even if the initial fee per unit time is equal to that for the elderly patient's visit.

For a provider paid a *salary* (or an institution receiving a *budget*), the financial opportunity cost of doing one thing over another is non-existent, as revenue is not related to what is done.[42, 43] However, if the new activity adds to the workload without generating more income, it represents a loss of leisure time for individuals or an increase in costs for an institution.

In a *capitated* environment, the opportunity cost is different again – every additional activity is an additional cost, and activities that may attract sicker patients or lead to greater subsequent activity will tend to be avoided, even in the face of a specific incentive, unless the marginal revenue from the incentive outweighs the longer run risk/cost. Therefore, it might be expected that incentives to undertake interventions that prevent complications in the near term (such as seasonal flu immunizations for older people) would be most readily accepted by a capitated provider, while incentives to undertake screening that might lead to identification of the need for further treatment (e.g., performing mammography) might be less effective. Of course, individual providers are rarely paid by capitation; therefore, as with salaried practice, the direct incentives upon the provider may be minimal or non-existent. Even where the provider's payment is based upon the unexpended share of capitation at the end of a period, this attenuates the incentive, since the capitation pool is usually shared across many providers, and, thus, the effects of an individual's practice on his or her payment may be small.

The specific incentive may also be influenced by other financial incentives in place. In addition, it may be related to the proportion of a provider's income that is dependent upon incentives other than the one being studied.[38] On the other hand, there is some evidence that providers do not vary practice style from patient to patient depending on insurance coverage but seem to adapt a style consistent with the dominant form of financial incentive.[44, 45]

Provider characteristics. Characteristics of the individual provider whose performance is being assessed might affect the impact the incentive has on quality. For example, the response to incentives might be expected to vary by provider age, gender, specialty, board-certification, country of graduation, whether full time or part-time, workload or total number of patients in panel, and proportion of patients/occasions of service per week where the incentive being studied is relevant.[18, 24, 38, 43, 46, 47]

In addition to these (relatively) easily observable factors, providers may differ in other ways that would be harder for a purchaser to assess but nonetheless could be important for response to an incentive. For instance, it is likely that the relationship between net additional income from an incentive and a provider's overall income and target income may influence the effectiveness of the incentive. A provider whose income is at or near a preferred income target may be less likely to respond to an incentive of a given amount than a provider who is not yet achieving his or her target income.[48]

A complete review of the many important psychological characteristics of individual providers that may influence the response to incentives—including intrinsic motivation, professionalism, and altruism[28-31] — is beyond the scope of this report. However, a forthcoming paper from Town et al. provides a valuable synthesis.[26]

Market characteristics. Characteristics of the market in which the provider is acting may also be important. For example, community-wide activities may increase provider cooperation and improve performance or lead to established norms—as the literature on small area variance has demonstrated.[49, 50] In addition, market factors such as managed care market share have been shown to influence provider practice patterns.[51] Since market-level phenomena change care, it is

conceivable they could also have an impact on a provider's predisposition to comply with a quality incentive.

Other predisposing factors. Other environmental factors may cause a provider to be more predisposed to accept and work to earn an incentive. These factors could include: trusting that the organization promoting the incentive has patients' and providers' bests interests in mind; believing performance measurement uses accurate, valid data; and having supportive medical leadership.[46]

Enabling Factors

Several factors may enable providers to respond more effectively when an incentive is offered. These may exist at the level of the organization in which the provider practices, or the patients that the provider sees. Enabling factors may also come from external sources—for instance when health plans adopt programs that facilitate providers' efforts to perform better.

Organizational characteristics. Organizational characteristics that may mediate the impact of an incentive on behavior include leadership, organizational culture, the organization of practice (partnership, company), size of practice, number of patients, and proportion of practitioners to whom the incentive is relevant.[18, 34, 38, 43, 44, 46] Other factors that may influence the impact of an incentive on quality are the use of electronic information systems for clinical data management, the implementation of guidelines related to the clinical focus of the incentive, utilization review,[52] peer pressure, educational activities,[53] and prior use of financial penalties for poor performance.[54]

Patient characteristics. Providers' responses to incentives may also be expected to vary according to characteristics of their patients, including purely clinical characteristics such as number of chronic conditions, but also age, gender, education level and insurance status and perhaps race and ethnicity.[18, 24, 43, 46, 55-57] For example, Irish general practitioners' responses to an incentive to limit their prescribing were found to vary according to the age of their patients.[58] In a randomized trial in the US in which physicians received clinical vignettes describing patients either as insured or uninsured, PCPs were more likely to recommend services to insured than to uninsured patients.[45]

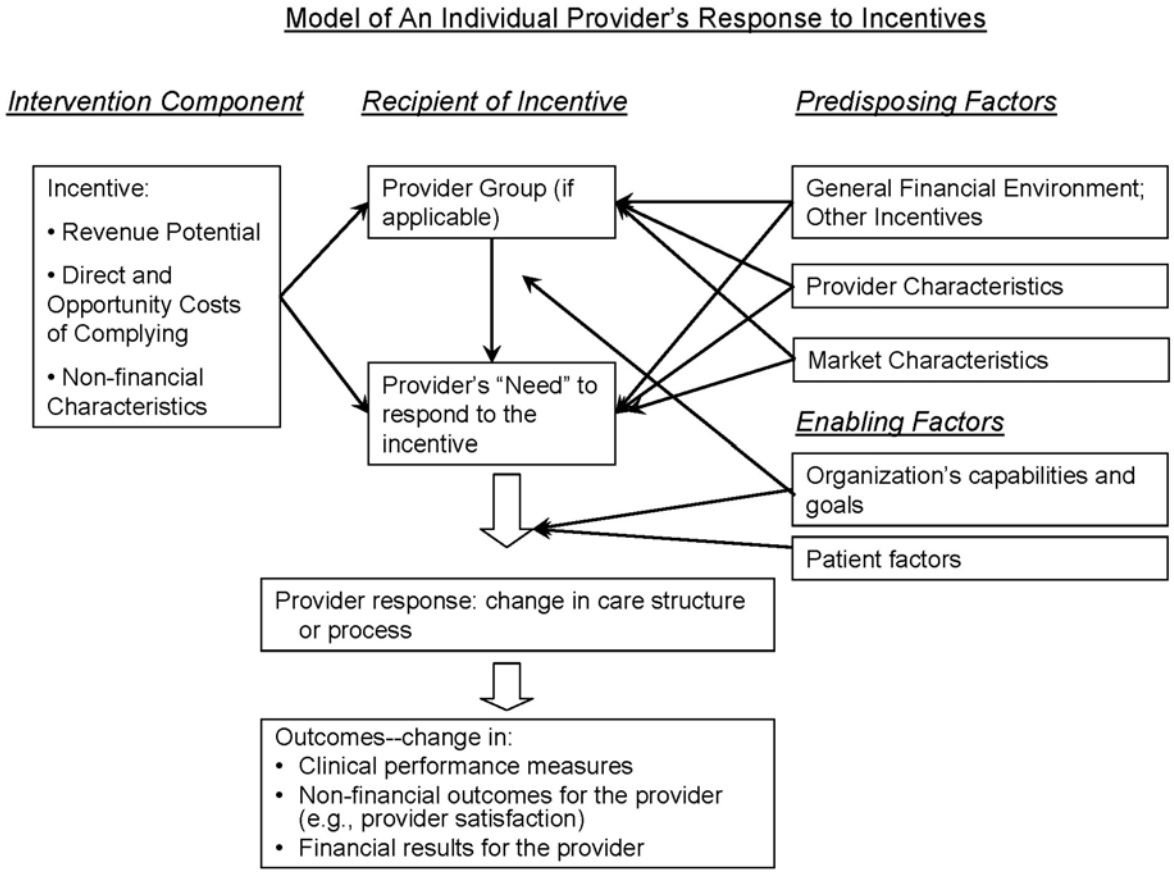
Other factors. Other factors may enable a provider to respond more effectively to an incentive. For instance, timely performance feedback from a health plan may facilitate providers' attempts to improve quality.[18, 32, 33, 35, 37, 38]

Conceptual Models of Individual Provider and Organizational Responses to Incentives

Drawing primarily on the health services research literature, but also on basic economic concepts that the health services literature does not address in research about specific incentives (e.g., the concept of opportunity costs), we propose the conceptual model in Figure 2 to understand the response of individual providers to incentives. In this model, we incorporate the six general determinants of physician behavior we describe above into a format that reflects Andersen's Behavioral Model.[25] Specifically, we propose that the financial and nonfinancial characteristics of an incentive are primary determinants of a provider's "need" to change practice in response to the incentive. This response, however, may be mediated by predisposing factors (e.g., the general

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

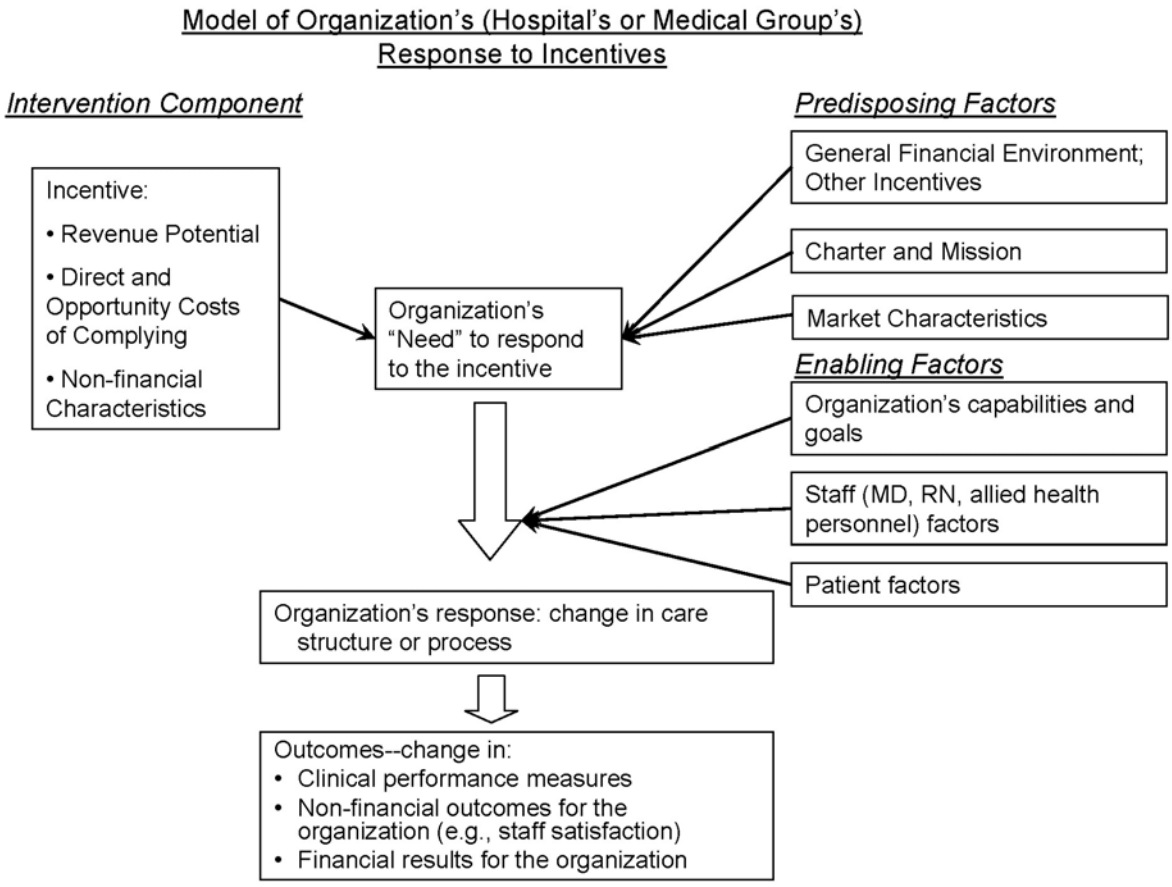
Figure 2: Model of an individual provider's response to incentives



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



2. Methods for Literature Search

Technical Expert Advisory Panel

For advice on the scope of the project, refinement of the key questions, and preparation of this technical review, we consulted technical experts in the following fields: employer purchasing strategies, provider performance assessment, consumer use of report cards and consumer preferences for health care information, risk adjustment, and economics. (See Appendix A, available at www.ahrq.gov/clinic/epcindex.htm.)

Target Audiences and Population

The decisionmakers addressed in this technical review are purchasers (both private purchasers such as employers and public purchasers such as the Centers for Medicare & Medicaid Services and State Medicaid programs), executives in health plans that must negotiate incentive arrangements with provider organizations or individual providers, executives in provider organizations that must negotiate incentive arrangements with providers, public health officials and other organizations interested in creating health care performance reports for public release, and policymakers. For the purpose of this report, provider organizations include all clinical health providers such as physicians, nurses, and hospitals. Public health officials and policymakers include those at the local, State, Federal, and international levels.

The ultimate target population of this report is the U.S. population at risk for morbidity or mortality resulting from quality problems in the provision of health care. We are interested in QBP strategies that affect the entire U.S. population—all members of which are at risk for receiving poor quality care—including those of all racial and ethnic backgrounds, all ages, and both genders.

Key Questions

We developed the key questions in collaboration with AHRQ, the Alliance (the nominating partner), and our Technical Expert Panel. The goal of these discussions was to identify the issues purchasers interested in QBP faced so that, if the available research offered conclusions about these aspects of QBP, the various stakeholders would be in a better position to select optimal approaches to QBP.

The key questions for which literature, ongoing research, or results from analyses were sought in preparation of this report were:

Choosing provider incentive strategies

1. What is the evidence on the extent to which health plans and employers use incentives to improve quality and efficiency?
2. Does the use of financial incentives for quality and efficiency actually increase the probability that patients receive high quality, efficient care?
3. Does the impact of financial incentives for quality and efficiency depend on:

- The basis of the incentive (structure, process, outcome)?
- The nature of the incentive (bonus, penalties or holdback, tiering or patient steerage/referral)?
- To whom the incentive is targeted (plan vs. provider group vs. individual provider)?
- The payer of the incentive (purchaser vs. plan vs. medical group)?
- The magnitude of the incentive?

4. Does the use of nonfinancial incentives for quality and efficiency actually increase the probability that patients receive high quality, efficient care?
5. Does the impact of nonfinancial incentives for quality and efficiency depend on:
    - The basis of the incentive (structure, process, outcome)?
    - The nature of the incentive (public release of performance report vs. confidential performance report)?

Relationship between cost and quality

6. Does greater spending result in higher quality?
7. What are the cost savings for the health care provider and purchaser as a result of the quality improvement?
8. What are the cost savings associated with different approaches to preventing medical errors or otherwise improving quality?
9. What specific processes and structures result in quantifiable cost savings? Who realizes the savings? How should they be shared?

Policy and market context in which incentives are used

10. What contextual variables (e.g., provider supply, employer number and market share, health plan competition, organizational system/infrastructure, employee demographics) positively or negatively influence the effectiveness of financial and nonfinancial incentives for providers?

Literature Review Methods

Based on input from our expert advisors, our conceptual model, and practical considerations, we developed literature review methods that included: inclusion and exclusion criteria to identify potentially relevant articles, search strategies to retrieve articles, abstract review protocols, and a system of scoring published studies for completeness.

Inclusion and Exclusion Criteria

To be considered an article that provided evidence regarding one of the key questions above, the article had to address one of the predictor variables and either quality (as measured by processes or outcomes) or cost. In addition, the intervention in the trial had to be a strategy that could plausibly be introduced by a purchaser. Our focus was on articles that provided definitive primary data from randomized, controlled trials, but we also included systematic reviews to

determine whether these contained any additional information not covered by the primary randomized, controlled trial reports.

We excluded articles that did not meet specific criteria in terms of the quality of the research and reporting. These were:

For interventional trials

- Intervention randomized
- Inclusion/exclusion criteria clear and appropriate
- Greater than 75% follow-up
- Note: two criteria usually used to judge the quality of a randomized, controlled trial—provision of placebo to the control group and blinding of the subjects—are not applicable in this situation

For systematic reviews

- Information source appropriate
- Information source adequately searched
- Inclusion/exclusion criteria clear and appropriate
- Data abstraction performed by at least 2 independent reviewers
- Principal measures of effect and the methods of combining results appropriate

Search Strategy

The objective of our search strategy was to identify all published QBP randomized trials and all ongoing research into QBP strategies. For the literature review, we used standard search strategies involving the querying of two online databases (MEDLINE® and Cochrane) using key words, followed by evaluation of the bibliographies of relevant articles, Web sites of relevant organizations (especially of funding agencies providing project summaries and of employer organizations pursuing QBP), and reference lists provided by our Technical Expert Panel (Table 1).

Table 1: Information sources for literature review and catalog of ongoing research

Goal of Search	Databases searched	Relevant Organizations (for Web-based searches)
Identify randomized, controlled trials of quality-based purchasing strategies	MEDLINE® Cochrane	AHRQ Robert Wood Johnson Foundation California HealthCare Foundation Commonwealth Fund National Business Coalition on Health Leapfrog Group

Database Searches

To identify potentially relevant articles in the medical literature, we searched MEDLINE® and Cochrane databases and references provided by our Expert Advisors.

MEDLINE® search strategies. We searched MEDLINE® (January 1980 to December 15, 2003) for English language articles using the search terms described in Table 2. Some citations were reviewed and articles were retrieved in more than one of the searches listed below.

Table 2: MEDLINE® searches to identify potentially relevant primary data

Search Terms	Citations reviewed	Articles retrieved
"pay" AND "quality" AND "measurement"	80	1
"incentive" AND "quality" AND "measurement"	195	5
"financial incentive" AND "quality" AND "efficiency"	125	11
"provider supply" AND "incentive"	15	0
"quality" AND "error" AND "safety" AND "cost*"	16	0
"pay" AND "performance"	389	2
"pay" AND "incentive" AND "quality"	79	3
"pay" AND "quality" AND "measurement" AND "Randomized Controlled Trial" [Publication Type]	8	1
"incentive" AND "quality" AND "measurement" AND "Randomized Controlled Trial" [Publication Type]	13	2
"financial incentive" AND "quality" AND "efficiency" AND "Randomized Controlled Trial" [Publication Type]	1	1
"provider supply" AND "incentive" AND "Randomized Controlled Trial" [Publication Type]	0	0
"quality" AND "error" AND "safety" AND "cost*" AND "Randomized Controlled Trial" [Publication Type]	0	0
"pay" AND "performance" AND "Randomized Controlled Trial" [Publication Type]	6	1
"pay" AND "incentive" AND "quality" AND "Randomized Controlled Trial" [Publication Type]	1	1
"incentive" AND "quality" AND "Randomized Controlled Trial" [Publication Type]	42	2
"pay" AND "quality" AND "Randomized Controlled Trial" [Publication Type]	26	2
"value" AND "incentive" AND "Randomized Controlled Trial" [Publication Type]	49	0
"value" AND "pay" AND "Randomized Controlled Trial" [Publication Type]	10	0
"Insurance, Health, Reimbursement" [MESH] AND "Randomized Controlled Trial" [Publication Type]	72	6
"Medicare Payment Advisory Commission" [MESH] AND "Randomized Controlled Trial" [Publication Type]	0	0
"Physician Payment Review Commission" [MESH] AND "Randomized Controlled Trial" [Publication Type]	0	0
"Prospective Payment Assessment Commission" [MESH] AND "Randomized Controlled Trial" [Publication Type]	1	0
"Prospective Payment System" [MESH] AND "Randomized Controlled Trial" [Publication Type]	28	1
"Salaries and Fringe Benefits" [MESH] AND "Randomized Controlled Trial" [Publication Type]	78	1
"Single-Payer System" [MESH] AND "Randomized Controlled Trial" [Publication Type]	2	0
"Fee-for-Service Plans" [MESH] AND "Randomized Controlled Trial" [Publication Type]	11	1
"Reimbursement Mechanisms" [MESH] AND "Randomized Controlled Trial" [Publication Type]	66	6

Search Terms	Citations reviewed	Articles retrieved
"Reimbursement, Incentive" [MESH] AND "Randomized Controlled Trial" [Publication Type]	10	4
"Cost and Cost Analysis" [MESH] AND "Randomized Controlled Trial" [Publication Type]	2,561	9
"Medical Errors" [MESH] AND "Randomized Controlled Trial" [Publication Type]	678	0
"Medication Errors" [MESH] AND "Randomized Controlled Trial" [Publication Type]	17	0
"Management Quality Circles" [MESH] AND "Randomized Controlled Trial" [Publication Type]	6	0
"Professional Review Organizations" [MESH] AND "Randomized Controlled Trial" [Publication Type]	3	0
"Quality Assurance, Health Care" [MESH] AND "Randomized Controlled Trial" [Publication Type]	586	14
"Quality Control" [MESH] AND "Randomized Controlled Trial" [Publication Type]	161	1
"Quality Indicators, Health Care" [MESH] AND "Randomized Controlled Trial" [Publication Type]	22	0
"Total Quality Management" [MESH] AND "Randomized Controlled Trial" [Publication Type]	45	2
"United States Agency for Healthcare Research and Quality" [MESH] AND "Randomized Controlled Trial" [Publication Type]	11	0
Total Articles	5413	76

The use of the asterisk expands search terms such that all combinations of terms with the phrase preceding the asterisk will be returned in the search (e.g., cost returns searches for cost, costs, etc.).
MESH = Medical Subject Heading

Cochrane search strategies. We searched the Cochrane databases from January 1, 1990 through December 15, 2003 (OVID, Evidence Based Medicine Reviews Multifile) using the search terms described in Table 3.

Table 3: Search terms and citations for Cochrane databases

Search terms	Citations reviewed	Articles retrieved
Pay	6	2
Incentive	4	0
Efficiency	74	0
Safety	264	0
Cost	210	2
Error	12	0
Performance	60	0
Value	95	0
Insurance	0	0
Reimbursement	0	0
Total	725	4

The use of the asterisk expands search terms such that all combinations of terms with the phrase preceding the asterisk will be returned in the search (e.g., cost returns searches for cost, costs, cost effectiveness, etc.).

Abstract Review

To identify potentially relevant articles for focused searching, at least two investigators (to ensure consistent application of the inclusion and exclusion criteria) reviewed each citation and, whenever an abstract was available, the abstract. Discrepancies in inclusion were resolved by discussion and re-review.

Evaluating Published Articles for Completeness of Reporting

We assessed each of the published articles for their completeness in reporting the factors we identified in our conceptual model that could influence a provider's response to incentives. Specifically, we scored them for the inclusion (or not) of descriptions of the elements in Table 4. We also recorded the type of care (preventive care, acute care, or chronic care) to which the quality measured pertained.

Table 4: Evaluating randomized controlled trials for completeness of reporting

Domain of the Conceptual Model	Specific Variable
Financial Characteristics of Incentive	*Recipient*: individual provider vs. provider group *Revenue potential*: magnitude of the financial incentive *Revenue potential*: incentive as a proportion of total income *Impact on cost*: direct costs and opportunity costs of complying
Nonfinancial Characteristics of Incentive	*Perceived attainability*: how easy/difficult it is to accomplish the task of the incentive *Performance domain measured*: structure, process, outcome
Predisposing Factors	*Financial characteristics of the environment*: proportion of income from: fee for service, salary, capitation *Financial characteristics of the environment*: number of other financial incentives in place *Provider characteristics*: demographics, specialty, and other immutable factors *Provider characteristics*: workload, proportion of patients if service where incentive relevant *Market characteristics*: community initiatives or performance standards
Enabling Factors	*Organizational characteristics*: size, type of practice, specialty, etc. *Organizational characteristics*: capabilities such as information systems, use of guidelines and feedback, etc. *Organizational characteristics*: leadership, culture, etc. *Patient characteristics*: demographics and other immutable factors *Patient characteristics*: type of insurance, benefits structure

Identifying Ongoing Research

Based on input from our expert advisors, our conceptual model, and practical considerations, we developed methods to catalog ongoing research into QBP that involved specifying: inclusion and exclusion criteria to identify potentially relevant research projects, search strategies to retrieve project abstracts, abstract review protocols, and a system of describing the study design of ongoing research projects.

Inclusion and Exclusion Criteria

Since the search for ongoing research focused on projects not yet reported in the literature, the criteria for identifying relevant projects focused on the planned intervention. Two types of research potentially met our inclusion criteria: projects designed as randomized controlled trials, or projects with interventions using QBP methods as described above (i.e., payment or performance reporting strategies) and applied at the community level (or in a broader geographic region, such as a State) that included historical or contemporaneous non-randomized control groups.

Search Strategy

We searched online health services research databases (HSRProj and AHRQ's Grants-On-Line Database or GOLD). We also searched the Web sites of other funders or coordinators of projects (e.g., the Leapfrog Group at www.leapfroggroup.org/RewardingResults/). Finally, we inquired of staff at AHRQ, the Robert Wood Johnson Foundation, the California HealthCare Foundation, and the Commonwealth Fund whether there was ongoing research that met our inclusion criteria being funded by those organizations. Table 5 lists our information sources for this aspect of the report.

Table 5: Information sources for the catalog of ongoing research

Goal of Search	Databases searched	Relevant Organizations (for Web-based searches and staff interviews)
Identify ongoing research evaluating quality-based purchasing strategies	GOLD (www.gold.ahrq.gov), HSRProj (via the National Library of Medicine at gateway.nlm.nih.gov/gw/Cmd)	AHRQ Leapfrog Group Robert Wood Johnson Foundation California HealthCare Foundation Commonwealth Fund

Database Searches

We searched the two available databases for ongoing health services research, using a similar search strategy for each (Tables 6 and 7). We accessed HSRProj through the National Library of Medicine's Gateway database at gateway.nlm.nih.gov/gw/Cmd and GOLD at www.gold.ahrq.gov.

GOLD search strategies. We searched GOLD through February 15, 2004 for grants funded by AHRQ using the categories described in Table 6. Through our combination of searches, we eventually evaluated all projects in GOLD.

Table 6: Search terms and citations for GOLD

Search by Category	Grants reviewed	Grants retrieved
Quality Outcomes	319	2
Quality Measures	189	2
Quality Improvement	256	2
Managed Care/Market Forces	98	1
Payment Strategies	22	1
Cost	121	0
New Knowledge	374	2
Total Grants	1379	10

HSRProj search strategies. We searched the HSRProj database through February 15, 2004 using the categories described in Table 7.

Table 7: Search terms and citations for HSRProj database

Search terms	Grant abstracts reviewed	Grants retrieved
Pay	49	1
Incentive	165	6
Efficiency	144	2
Safety	374	4
Error	160	1
Performance	546	7
Value	219	6
Reimbursement	136	2
Total Grants	1793	29

The use of the asterisk expands search terms such that all combinations of terms with the phrase preceding the asterisk will be returned in the search (e.g., cost returns searches for cost, costs, cost effectiveness, etc.).

Grant Abstract Review

Two investigators reviewed the abstracts of projects identified from the database searches to assess relevance to the technical review. Discrepancies in inclusion were resolved by discussion and re-review and by discussion with project officers at funding agencies or with the principal investigator of the project under consideration.

Describing the Study Design of Ongoing Research

For each research project, we interviewed either project staff (usually the principal investigator) or the project officer to determine the study design. We obtained information about the intervention—performance measures and incentives used—and the control group. The information sought is described in Table 8.

Table 8: Design information sought about ongoing research

Design Issue	Examples of Possible Responses
Patient Population from an Insurance Perspective	Privately Insured, Medicare, Medicaid, or multiple populations
Health Plan Setting	Health maintenance organization, preferred provider organization, point of service
Control Group	Randomized controlled trial vs. non-randomly selected contemporaneous control vs. historical control
Incentive Structure	Describe financial or reputational gains from superior performance
Performance Measures	Participation vs. clinical performance (for the latter, describe determinants of performance assessment, including weighting given when multiple measures are used)
Evaluation Plan/Goals	Assess determinants of participation in the program, catalog incentives used, test impact of incentives on clinical performance

3. Results for Literature Search

This chapter presents the results of our systematic review of the literature, our search for ongoing research, and our evaluations of outcomes reports.

Synthesis of Literature About Quality-based Purchasing

Articles Identified

Our literature searches identified 5,045 unique candidate articles for inclusion in our literature review (Figure 4). Of these, 4,882 were eliminated after review of their abstracts. The reasons for exclusion were: 4,861 because they were not relevant to the key questions, 14 because they were cost effectiveness studies or decision analyses that provided no primary data about the questions, and 7 because they had dependent variables that were "quality" in an abstract sense—responses to a questionnaire or survey about what the provider would do if presented with a hypothetical patient—rather than actual measurement of quality performance.

The remaining 163 articles underwent full text review, which eliminated another 101 that were not relevant to the study question. Of the 62 studies that were relevant, only 15 were good quality. Of these, 9 were interventional studies (randomized controlled trials)[18, 32-34, 37, 38, 43, 59-61] and 6 were systematic reviews (Figure 4).[20, 46, 62-65] Of the nine randomized controlled trials, eight used specific financial incentives as the intervention;[18, 32-34, 37, 38, 43, 59, 60] one used specific reputational incentives as the intervention.[61]

Completeness of Reports of Randomized Controlled Trials of Incentives

Trials of specific financial incentives. In every article reporting the results of a randomized controlled trial of performance-based payment incentives, there were significant variables from our conceptual model that were either not reported at all or that were incompletely described. In Table 9 we show the completeness in the reporting of the eight trials of specific performance-based payment.[18, 32-34, 37, 38, 43, 59, 60]

The only variables that were reported in all trials were characteristics of the incentive itself: the recipient (although even this was sometimes ambiguous between individual provider versus provider group), the magnitude of the incentive, and the domain of performance measured. Several potentially critical variables were never reported in any trial, including payment incentive as a proportion of total income and the costs of complying with the incentive and most enabling factors at the organizational level.

Figure 4: Articles Identified By Systematic Searches

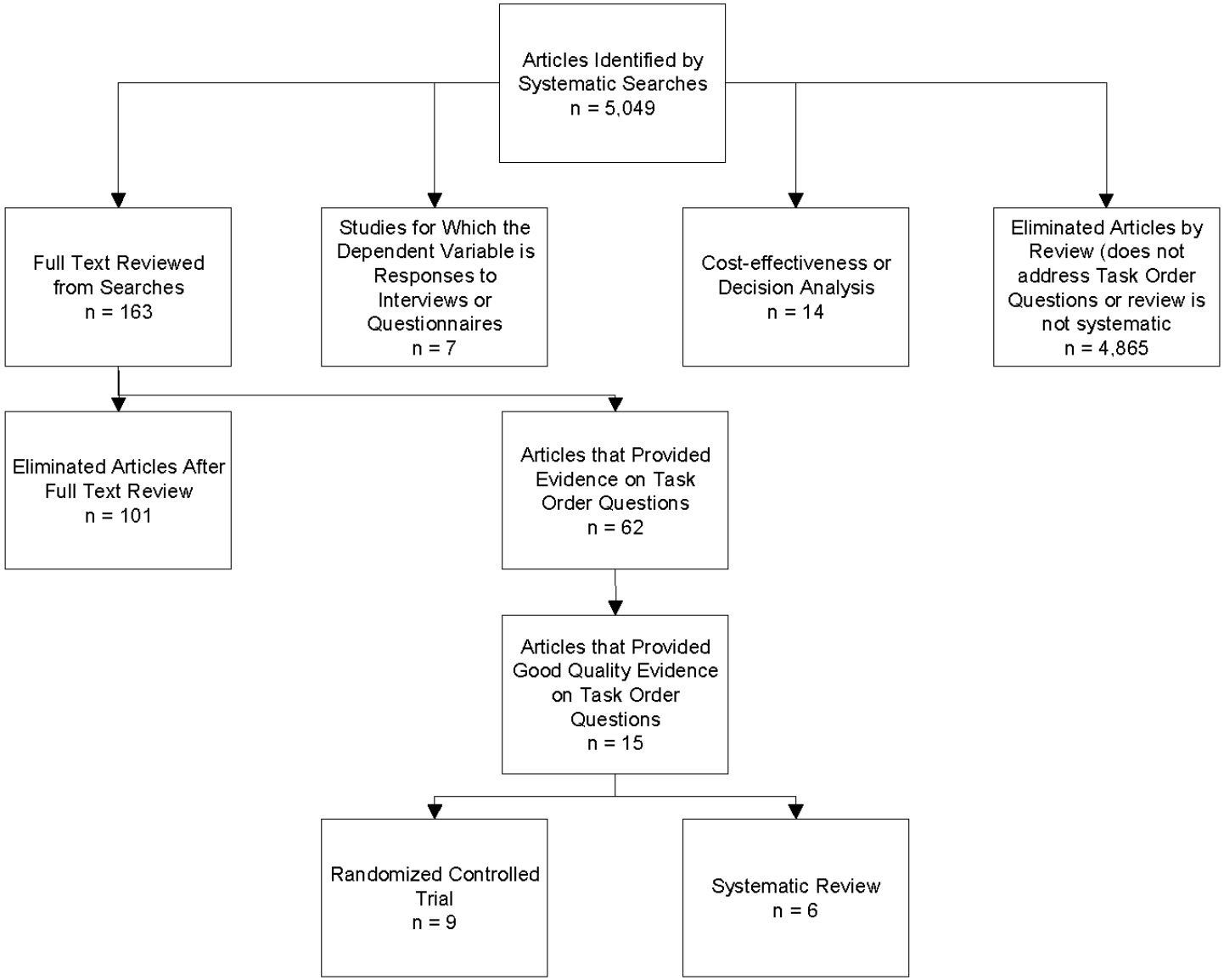


Table 9. Evaluating Randomized, Controlled Trials for Completeness of Reporting

Domain of the Conceptual Model	Specific Variable	Christensen	Davidson	Fairbrother	Hickson	Hillman '98	Hillman '99	Kouides	Roski
Financial Characteristics of Incentive	*Recipient of the incentive*: individual provider vs. group	Reported (individual pharmacist)	Reported (individual physician)	Reported (individual physician)	Reported (individual physician)	Reported (individual physician or group)	Reported (medical group)	Reported (individual physician or group)	Reported (individual physician)
	Revenue potential: magnitude of the financial incentive	Reported (schedule of fees-for-service)	Reported (schedule of fees-for-service)	Reported (bonus up to $7,500 vs. fees-for-service vs. control)	Reported ($2/visit fee-for-service)	Reported (~33% chance to receive a bonus up to $5,000)	Reported (~10% chance to receive a bonus; total potential $ not reported)	Reported (fee-for-service, $0.80-$1.60 per vaccination)	Reported (bonus up to $10,000)
	Revenue potential: incentive as a proportion of total income	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
	Impact on cost: direct costs and opportunity costs of complying	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
Non-financial Characteristics of Incentive	*Perceived attainability*: How easy/difficult it is to accomplish the task of the incentive	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
	Performance domain measured: structure, process, outcome	Reported (chronic care process: medication instruction)	Reported (preventive care process: well child continuity visits)	Reported (preventive care process: vaccinations)	Reported (preventive care process: well child visits)	Reported (preventive care process: vaccinations)	Reported (preventive care process: cancer screening)	Reported (preventive care process: vaccinations)	Reported (preventive care process: tobacco screening, tobacco cessation)

Table 9. Evaluating Randomized, Controlled Trials for Completeness of Reporting (cont'd)

Domain of the Conceptual Model	Specific Variable	Christensen	Davidson	Fairbrother	Hickson	Hillman '98	Hillman '99	Kouides	Roski
Predisposing Factors	*Financial characteristics of the environment*: proportion of income from: fee for service, salary, capitation	Not reported	Not reported	Report that all apply, but do not give percentages	Reported (salary)	Not reported	Not reported	Not reported	Report that all apply, but do not give percentages
	Financial characteristics of the environment: number of other financial incentives in place	Not reported	Not reported	Not reported	Not reported	Report many other incentives, but do not describe them	Report many other incentives, but do not describe them	Not reported	Not reported
	Provider characteristics: demographics, specialty, and other immutable factors	Not reported	Not reported	Only board certification reported	Only specialty reported	Only specialty reported	Only specialty reported	Not reported	Specialty reported
	Provider characteristics: workload, proportion of patients where incentive is relevant	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
	Market characteristics: community initiatives or performance standards	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported

Table 9. Evaluating Randomized, Controlled Trials for Completeness of Reporting (cont'd)

Domain of the Conceptual Model	Specific Variable	Christensen	Davidson	Fairbrother	Hickson	Hillman '98	Hillman '99	Kouides	Roski
Enabling Factors	*Organizational characteristics*: size, type of practice, specialty, etc.	Not reported	Not reported	Not reported	Reported (type and specialty)	Reported (varies)	Reported (varies)	Size reported	Reported
	Organizational characteristics: capabilities such as information systems, use of guidelines and feedback, etc.	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
	Organizational characteristics: leadership, culture, etc.	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
	Patient characteristics: demographics and other immutable factors	Not reported	Not reported	Age, high poverty levels reported	Age, high poverty levels reported	Age and race reported	Age reported	Age reported	Not reported
	Patient characteristics: type of insurance, benefits structure	Reported (Medicaid patients, no cost to patients)	Medicaid patients, benefits structure not reported	Not reported	Reported (most uninsured)	Medicaid patients, benefits structure not reported	Medicaid patients, benefits structure not reported	Medicare demonstration project patients, benefits structure not reported	Not reported

A trial of reputational incentives. There was a single trial of reputational incentives, in which Hibbard et al. report on the response of hospitals in south central Wisconsin to public release of performance data and compare this response to other Wisconsin hospitals who were randomly assigned to receive either a confidential report or no report at all.[61] We did not include this study in Table 9 because some of the elements of that table are not applicable to or not measurable for reputational incentives (e.g., most of the financial characteristics variables) and others were not applicable to the specifics of the study (e.g., market characteristics will vary when a study is done statewide). In this article, however, there was no explicit consideration of whether response to the incentive varied with differences among hospitals in terms of enabling or predisposing factors, which were not measured.

Results of Randomized Controlled Trials of Performance-based Payment

The eight trials of performance-based payment were neither consistent in their design of the independent variable (the financial incentive offered) nor comparable in terms of their dependent variable (the performance indicator measured). Thus, we present their results as a function of several of the variables within the conceptual model (those that are actually reported for all papers). Note that among these eight trials there were ten hypotheses tested, because one study had two intervention arms (a fee-for-service arm and a bonus arm) compared to controls[37] and one had two dependent variables (smoking cessation processes and smoking cessation outcomes).[60]

Recipient of incentive. In four studies, the recipient of the incentive was an individual provider,[32, 33, 37, 43, 59, 60] while in the other four the recipient was the provider group or could be either an individual provider or a group.[18, 34, 38] Among the studies targeting individual providers, there were five positive and two negative results; among the studies in which the target was always or could be the provider group, there were one positive and two negative results. (In general, we use the term "positive" to mean an effect in the desired direction—the incentive worked—and "negative" to mean there was no significant effect of the incentive on the outcome measure.)

In seven studies, with a total of nine dependent variables, the target of the incentive was a physician. Of the nine dependent variables assessed, five showed a significant relationship to the incentive in the expected direction, four showed no significant change after the incentive was introduced.[18, 34, 37, 38, 43, 59, 60] A single study (reported in two papers) involved non-physician recipients (pharmacists) and was positive.[32, 33]

Magnitude of the incentive. Incentives ranged in magnitude from $0.80/flu shot[34] to a bonus of up to $10,000 per clinic per year.[60] There was no consistent relationship between the magnitude of the incentive and response, and in fact the largest single incentive (the bonus of up to $10,000) was ineffective.[60] The two studies in which the provider faced significant uncertainty about whether they could achieve success—in each case because the incentive was tied to performance *relative* to other groups, and this benchmark was unknown during the time when performance was measured—were negative.[18, 38]

Structure of the incentive. Five studies (with five outcomes) assessed fee-for-service incentives to improve quality,[32-34, 37, 43, 59] while four studies (with five outcomes) evaluated the impact of bonuses tied to performance.[18, 37, 42, 60] Among the studies of fee-for-service, four were

positive and one was negative. With bonuses tied to performance, there were two positive results and three negative.

Performance domain measured. Among the articles included, there were seven studies of preventive care with nine dependent variables assessed. Among these nine outcomes, five were positive and four negative. The single study addressing chronic care was positive.[61]

Patient factors. Authors did not report the burden adherence would place on patients in any of the articles we found. However, in a general sense, we found that incentives to achieve performance were more effective when the indicator to be followed required less patient cooperation (e.g., receiving vaccinations or answering questions about smoking) than when significant patient cooperation was needed (e.g., to quit smoking, Table 10).

Table 10: Available results by conceptual model domains tested

Conceptual Domain and Specific Variable	Results
Financial Characteristics of the Incentive: Recipient Individual vs. Group	• Individual: 5 positive, 2 negative • Group or Individual: 1 positive, 2 negative
Financial Characteristics of the Incentive: Recipient Provider Type	• Physicians: 5 positive, 4 negative • Pharmacists: 1 positive
Financial Characteristics of the Incentive: Magnitude	• No clear relationship between magnitude and result • Both trials in which the performance required to achieve a bonus was unknown were negative
Nonfinancial Characteristics of the Incentive: Performance Domain Measured	• Preventive care: 5 positive (3 immunizations, 1 well-child, 1 tobacco screening); 4 negative (1 cancer screening, 1 well-child, 1 immunizations, 1 tobacco cessation) • Chronic care: 1 positive
Patient Factors	• Goals likely to encounter fewer patient barriers (immunizations, tobacco screening): mostly positive • Goals that required modest patient cooperation (e.g., well child visits and cancer screening): mixed • Goals that require significant patient cooperation (e.g., tobacco cessation): negative

Synopses of the available studies. As there were so few available studies, we are able to include synopses of each in this report. Rather than use the original abstracts, which varied in structure and content, we have put each into a uniform format. The eight randomized controlled trials of performance-based payment, presented in alphabetical order by first author, were:

- **Christensen DB, Holmes G, Fassett WE, et al. Influence of a financial incentive on cognitive services: CARE project design/implementation. *J Am Pharm Assoc.* Sep-Oct 1999;39(5):629-639**.

  and

- **Christensen DB, Hansen RW. Characteristics of pharmacies and pharmacists associated with the provision of cognitive services in the community setting. *J Am Pharm Assoc.* Sep-Oct 1999;39(5):640-649.**

Setting and Design: This study took place in Washington State from February 1994 – September 1995. Incentives were offered by the Washington State Cognitive Activities and Reimbursement Effectiveness Project to community pharmacies that served primarily

ambulatory patients, were not a part of a staff-model health maintenance organization, and dispensed at least 50 prescriptions per month to ambulatory Medicaid recipients to improve performance. The treatment group (n=110) performed and documented cognitive services (CS) provided to Medicaid recipients, received a fee for each intervention of $4 or $6 dollars depending on whether the CS lasted greater than six minutes, and received a monthly stipend of $40/month for their participation in the demonstration. Control pharmacies (n=90) received a monthly participation stipend of $40/month, but performed and documented CS interventions without additional reimbursement. A silent control group (group C, n=100) neither received additional payment nor documented CS interventions. Performance was measured over 19 months, ensuring a minimum 12 month observation period for each pharmacy.

Results: At baseline, differences in operating characteristics between groups were minor and nonsignificant. Over the study period, the incentive group performed significantly more CS than the control group. Factors associated with the provision of any CS by pharmacists included perceptions of how burdensome the task of documenting CS was and the percentage of sales from prescriptions.

- **Davidson SM, Manheim LM, Werner SM, Hohlen MM, Yudkowsky BK, Fleming GV. Prepayment with office-based physicians in publicly funded programs: results from the Children's Medicaid Program. *Pediatrics.* Apr 1992;89(4 Pt 2):761-767.**

Setting and Design: This study took place in Suffolk County, New York, from July 1983 through December 1985. The Health Care Financing Administration and the John A. Hartford Foundation offered incentives to individual primary care physicians in private office based practices. All 140 primary care physicians who treated Medicaid children and had more than $2000/year in billings were invited to participate and 80 agreed. Physicians were randomly assigned to augmented fee-for-service (n=40) at nearly double the usual New York Medicaid rates in return a commitment to meet performance goals or capitation (n=40) and compared to physicians operating under conventional Medicaid arrangements. (We do not report on the capitation arm herein as comparisons of capitation to fee-for-service were not within the scope of this report.) The payment groups were evaluated in comparison to children enrolled in the regular Children's Medicaid Program and the patients who refused to be included in the study to see if there was any difference between the groups. Performance was measured over 29 months.

Results: There was no difference in the rates of compliance with well-child care recommendations between the augmented fee-for-service group and the control group. Emergency room visit rates and hospitalization rates also were not significantly different.

- **Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. *Ambul Pediatr.* 2001;1(4):206-212.**

Setting and Design: This study took place in New York City, NY from July 1997 to December 1998. Incentives were offered to individual inner-city physicians with the highest rates of poverty and proportions of Medicaid-enrolled children among their patients to

determine the effect of two financial incentives—bonus and enhanced fee-for-service—on documented immunization rates during a second period of observation. Physicians assigned to the bonus with feedback group (n=24) could receive $1000 and $2500 for improvements in immunization rates of 30% and 45% from baseline, $5000 for reaching 80% up to date (UTD) coverage, and $7500 for reaching 90% UTD coverage for immunizations against diphtheria and tetanus toxoids and pertussis vaccine (DTP), *Haemophilus influenzae* type b vaccine (Hib), polio vaccine, and rubella vaccine (MMR). The investigators also determined the percentage of visits in the past four months that were missed opportunities to immunize (MOI) and to increase the average number of vaccinations per child given on a date of visit versus no office visit scheduled. Physicians assigned to the enhanced fee-for-service group (EFF, n=12) received $5/vaccine administered within 30 days of coming due and $15/visit at which all vaccines were administered. The control group (n=21) received feedback on their performance with respect to lead, anemia and overall UTD screening and $100 for participation in a concluding interview. The incentives were given in 4-month intervals. Performance was measured over 16 months.

Results: Overall UTD coverage increased in the two groups receiving financial incentives. UTD coverage improved significantly within the bonus group compared to the control between time 1 and time 3, and in the EFF group at time 4. The average number of immunizations recorded in the chart increased significantly for children in the bonus group between time 1 and time 2, but not for children in the EFF group relative to the control. The MOI for sick visits were high, ranging from 89-92% and did not change significantly for the EFF group, whereas they decreased significantly at time 3 for the bonus group relative to the control. Seventy-one percent of the visits were sick visits, thus a change in this category will have an overall effect.

- **Hickson GB, Altemeier WA, Perrin JM. Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study. *Pediatrics.* Sep 1987;80(3):344-350.**

Setting and Design: This study took place at the Vanderbilt University Pediatric Residents Continuity Clinic in Nashville, Tennessee from September 1983 to June 1994. Incentives were offered by the study to 18 medical residents. Nine were randomized to receive $2/patient visit and nine were randomized to a control group that received the expected average compensation of $20/month to determine the effect of augmented fee-for-service on physician behavior. Prior to data collection residents completed a questionnaire to monitor interest in outpatient practice and a variety of other questions. Performance was measured over nine months.

Results: Due to the small sample size, randomization failed to equalize physician interest because the nine physicians in the control group were more likely to plan a career in private practice than the fee-for-service group. Fee-for-service physicians did not have significantly more patient visits, but fee-for-service patients experienced greater continuity of care (more often saw their regular physician when they came to clinic) and fewer ER visits than patients enrolled to salaried physicians. There were 22% more per capita visits by patients using fee-for-service than by patients with control physicians, almost entirely due to well-child visits.

Although initial and follow-up visits for illness were not different, fee-for-service patients averaged 43% more well child visits.

- **Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. *Am J Public Health.* Nov 1998;88(11):1699-1701.**

Setting and Design: This study took place in Philadelphia, PA in 1993-1995. Incentives were offered by a Medicaid managed care organization structured like an independent practice association with provider sites paid by capitation, to the largest primary care sites stratified by practice type (solo/group) to ensure sufficient representation of each. The randomly selected intervention sites (n=26) were eligible to receive a full bonus (20% of capitation) for the three intervention sites with the highest compliance scores, the three next highest scores and the three improving the most from the previous audit both received partial bonuses (10% of capitation). The in order to increase their rates of compliance in mammography, Pap smear, and colorectal screening for all female members fifty years of age and older. In addition to bonuses, the intervention group received feedback. The control group (n=26) received no intervention and no feedback. Bonuses ranged from $570 to $1260 per site. Chart audits were performed at baseline and every six months for 1.5 years

Results: There was no significant difference between intervention and control groups by type of practice, specialty, or patient panel size. Baseline compliance scores were relatively low and did not differ significantly between study groups, although group practices had consistently higher compliance scores than solo practices. There was a significant improvement over time in performance for both intervention and control groups, but there was no significant difference between the groups. A subanalysis comparing aware and unaware intervention sites showed no significant between-group differences.

- **Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. *Pediatrics.* Oct 1999;104(4 Pt 1):931-935.**

Setting and Design: This study took place in Philadelphia, PA in 1993-1995. Incentives were offered by a Medicaid managed care organization structured like an IPA with provider sites paid by capitation, to primary care physicians with at least twenty-five pediatric members younger than seven. After stratification by practice type (solo/group), the primary care sites were randomly assigned into one of three groups to assess whether feedback coupled with financial incentives could improve pediatric preventative care. The three arms included a feedback only group (n=17) where physicians received written feedback about compliance scores, a feedback and incentive group (n=19) where physicians received feedback and a financial bonus when compliance criteria were met, and a control group (n=17) with no feedback and no incentive. Preventive care guidelines were distributed to providers in all three study groups. Chart audits were performed for practice sites in all three groups at 6-month intervals. Eligibility for bonuses in the feedback and financial incentives group was based on a total compliance score of 20% for each indicator. The three sites with the highest total compliance received a full bonus (20% of the sites total 6-month capitation

for pediatric members less than seven years of age). The three next best scoring sites received a partial bonus (10% of the sites total 6-month capitation for pediatric members less than seven years of age). The three sites showing the most improvement from the last audit the partial bonus if their total compliance score increased by at least 10%. Performance was measured at baseline and every six months for 1.5 years.

Results: Bonuses paid out during the course of the study ranged from $772-$4682 with an average of $2000. Thirteen of nineteen sites received a bonus. At baseline no significant differences were observed. Compliance with pediatric preventive care improved dramatically in the study period. Repeated measures analysis of variance demonstrated a significant increase in all three study groups throughout the time in total compliance (56%-73%), as well as scores for immunizations (62%-79%) and other preventive care (54%-71%). However no significant differences were observed between the intervention groups and the control, nor were there any interaction (group-by-time) effects.

- **Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. Performance-based physician reimbursement and influenza immunization rates in the elderly. The Primary-Care Physicians of Monroe County. *Am J Prev Med.* Feb 1998;14(2):89-95.**

Setting and Design: This study took place in Rochester, New York and surrounding Monroe County from September 1991 to January 1992. Incentives were offered by the Medicare Influenza Vaccine Demonstration Project to providers or group practices who provided primary care to at least fifty patients sixty-five years and older, participated in the Medicare Demonstration Project, and used target-based poster method for tracking immunizations. Physicians were randomized by practice group to the control (n=27) or the incentive group (n=27), which was eligible for reimbursement above the standard $8 fee per immunization if immunization rates above 70% or 85% were achieved. If a final immunization rate of 70% was attained, the physician received an additional 10% reimbursement—$.80/shot given in the office. If a final immunization rate of 85% was attained, the physician received an additional 20% reimbursement—$1.60/shot given in the office. Immunizations given outside the office were included in the percent immunized, but were not given the incentive. Performance was measured over three months.

Results: At baseline there were no statistically significant differences between the control and incentive groups. The median change in immunization rate was significant (10.3%) in the incentive group and not significant (3.5%) in the control group. In the incentive group, 52% of practices attained the 70% immunization target level, with 15% attaining the target level of 85%. In the control group, 44% of practices attained the 70% immunization target level, with 7% attaining the target level of 85%. Individual physician performance within group practices was quite variable.

- **Roski J, Jeddeloh R, An L, et al. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. *Prev Med.* Mar 2003;36(3):291-299.**

Setting and Design: This study took place in the Minnesota from May 1999 to December 2000. Incentives were offered by the Allina Health System to forty clinics providing primary care service (family practice, internal medicine, obstetrics/gynecology) in a large multispecialty group practice to improve performance. The three experimental conditions were represented by financial incentives for reaching preset clinical performance targets combined with access to a centralized smoker registry and intervention system (Incentive + Registry group, n=10), financial incentives for reaching preset clinical performance targets (Incentive group, n=15), and no intervention except the distribution of printed versions of the smoking cessation guidelines (Control group, n=15). The two clinical performance targets were 75% of adult patients having their tobacco status clearly identified at each visit and documented in their medical records and 65% of smokers having received ongoing in-office counseling (measured as advice to quit given at last visit). Actual smoking cessation rates were a secondary endpoint. Incentive amounts were based on the number of providers per clinic. Clinics with one to seven providers could receive $5000 and clinics with eight or more providers were eligible for a $10000 bonus. Clinics who reached or exceeded only one of the two performance goals were eligible for half the incentive. The Incentive + Registry group received weekly updates on their referral activity during the past week and their referral activity to date. The Incentive + Registry group was able to compare the referral patterns of their site to other clinics. Performance was measured over nineteen months.

Results: At baseline no differences were found between the groups. Identification of patients' tobacco use status statistically significantly improved in all groups but was statistically significantly higher in the two incentive groups (14.4% in the Incentive group and 8.1% in the Incentive + Registry group vs. 6.2% in the control group). However, ongoing in-office counseling and actual quit rates did not differ significantly between the groups.

Results of Randomized Controlled Trials of Reputational Incentives

There was only one randomized controlled trial of reputational incentives.[61] This study showed that hospitals with low performance scores were more likely to engage in quality improvement activities. This was especially true for hospitals whose performance was released to the public (as opposed to being kept confidential). As this is the only study of this type, we include a synopsis of it below:

- **Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood).* Mar-Apr 2003;22(2):84-94.**

Setting and Design: This study took place in Wisconsin and concluded in May 2002. The study evaluates the impact of a public hospital performance report on subsequent hospital quality improvement efforts. The report on hospital safety was produced and disseminated by the Alliance, a large employer-purchasing cooperative in the Madison, Wisconsin area. The report, Quality Counts, compared the performance of twenty-four hospitals in south

central Wisconsin. Two summary indices of adverse events (deaths and complications) occurring within the broad categories of surgery and non-surgery were included, along with indices in three individual clinical areas: hip/knee surgery, cardiac care, maternity care. Hospitals were rated as better than expected (fewer deaths and complications), as expected, or worse than expected. The primary intervention group was the twenty-four hospitals in south central Wisconsin in the Alliance service area. These hospitals were in the public report, were not randomly selected, and received a more detailed report on their performance. The other ninety-eight hospitals in Wisconsin were randomly assigned to either the secondary intervention that received a private report on their performance or the control condition that did not receive anything.

Results: On average, public, private, and no report hospitals were slightly negative about the idea of publicizing hospital performance. There were statistically significant differences among the respondents toward the validity of the public report, its appropriateness for the public's use, and its value for quality improvement. Public report hospitals were most negative and those with private reports were most positive. Low-scoring public-report hospitals show the highest level of quality improvement activities, the private-report hospitals an intermediate level, and the no report hospitals the lowest level and the differences among the hospitals in the three study conditions were statistically significant. Most of the hospitals were optimistic that they could improve their scores through attention to quality improvement within the next two years.

Ongoing Research Into Quality-based Purchasing

Ongoing Randomized Controlled Trials

We identified no currently ongoing randomized controlled trials of QBP strategies from any funding source.

Interventional Trials With Non-Randomized Designs

There were 18 ongoing research projects in which there was a QBP intervention without randomization (Tables 11 and 12). For many of these, the exact nature of the performance measures and the incentive were still being determined. For some, the study design is observational; that is, health plans are making decisions about incentives without input from the investigators, but the investigators are assessing the response.

The single largest initiative is Rewarding Results, which has components funded by the Robert Wood Johnson Foundation, the California HealthCare Foundation, and AHRQ. Therefore, we first list the Rewarding Results projects (Table 11), then list separately other ongoing QBP research (Table 12).

Topics covered. These projects will provide some important additional information about QBP. Several studies (particularly those by Rabson et al. and Epstein et al.) will describe the type and frequency of use of QBP strategies. Several projects (most of the Rewarding Results projects, plus Young et al., Braun et al. and Callahan et al.) will investigate provider reactions to

incentives in terms of willingness to participate in programs and awareness of the incentives offered. In addition, Braun et al. and Young et al. will obtain quantitative and qualitative information about attitudes towards incentives used and performance targets set (such as salience, clinical validity, and whether the performance measures were within the providers' scope of control). These studies may be useful for understanding providers' motivation to respond and organizational decisionmaking when incentives are offered. Still other projects (particularly Sofaer et al.) will report on the tools used to communicate incentives, rather than the provider or consumer response to the incentive.

Quantitative assessments of the impact of incentive interventions. The Rewarding Results projects and several others (Delbanco et al., Rosenthal et al., and Epstein et al.) will provide assessments of the impact of incentives on traditional performance measures of structure, process, and outcomes. While none of these are randomized and all involve organizations that self-select to adopt or participate in incentive programs, taken together they will provide preliminary evaluations of QBP in Medicaid, Medicare, and commercial insurance settings and will cover many different approaches to incentives. For instance, in the Integrated Healthcare Association project alone, the health plans have adopted financial incentives that vary in structure from increases in capitation to augmented payments per encounter (and also range widely within these approaches; e.g., there is greater than two-fold variation in the magnitude of the capitation increase available across plans). One of these studies (Rosenthal et al.'s "Determining Whether Pay-for-Performance Incentives Improve Health Care Quality in Medical Groups") will investigate whether the provision of incentives for specific indicators also lead to improvement in domains of performance that are not included in the incentive measure set (or, alternatively, worsening in these measures if the non-incentivized areas of performance are subsequently neglected). In addition to these that are ongoing, the Centers for Medicare & Medicaid Services has a project in development with Premier Healthcare Informatics that will include both financial and reputational incentives for hospitals. While the dissemination of performance data has already begun, the evaluation plan for this project is still under development (see: www.cms.hhs.gov/researchers/demos/phqidemo.asp).

Among the interventional studies, there are also some major differences in the characteristics of the incentives themselves between the prior literature and the ongoing research. For instance, the ongoing studies involve actual health plans or government programs making an ongoing commitment to an incentive strategy, rather than a researcher making a short-term payment intervention (which was the situation in the prior studies). Similarly, all the studies included in the literature review above involved incentives directed at only a small number (usually just one) performance indicator for a single condition or type of patient. However, all the ongoing interventional studies we identified involve multiple measures (often ten or more) across a variety of conditions and distinct patient populations. Both these factors—that the incentive comes from a health plan or government program that expresses a longer-term commitment to the strategy and that there are multiple indicators—may should increase the probability that providers will believe that investments in quality improvement (such as installing a new information system) can be recouped relative to previously studied incentive strategies.

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.)

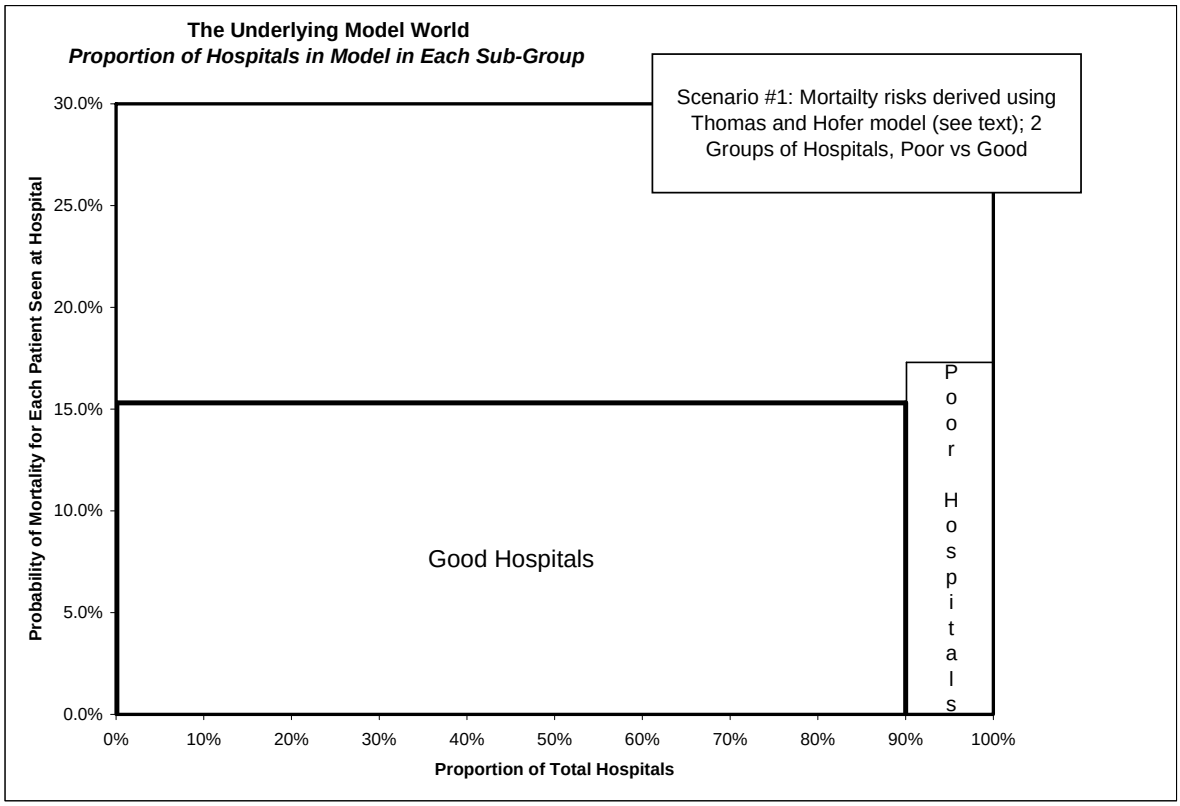
Given the assumptions made in each scenario, we then apply a performance label to each individual hospital (e.g., "poor" vs. "good" vs. "superior"). We refer to this labeling process as the *evaluation system*, and the frequency of mislabeling is determined both by the assumptions about the hypothetical 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.

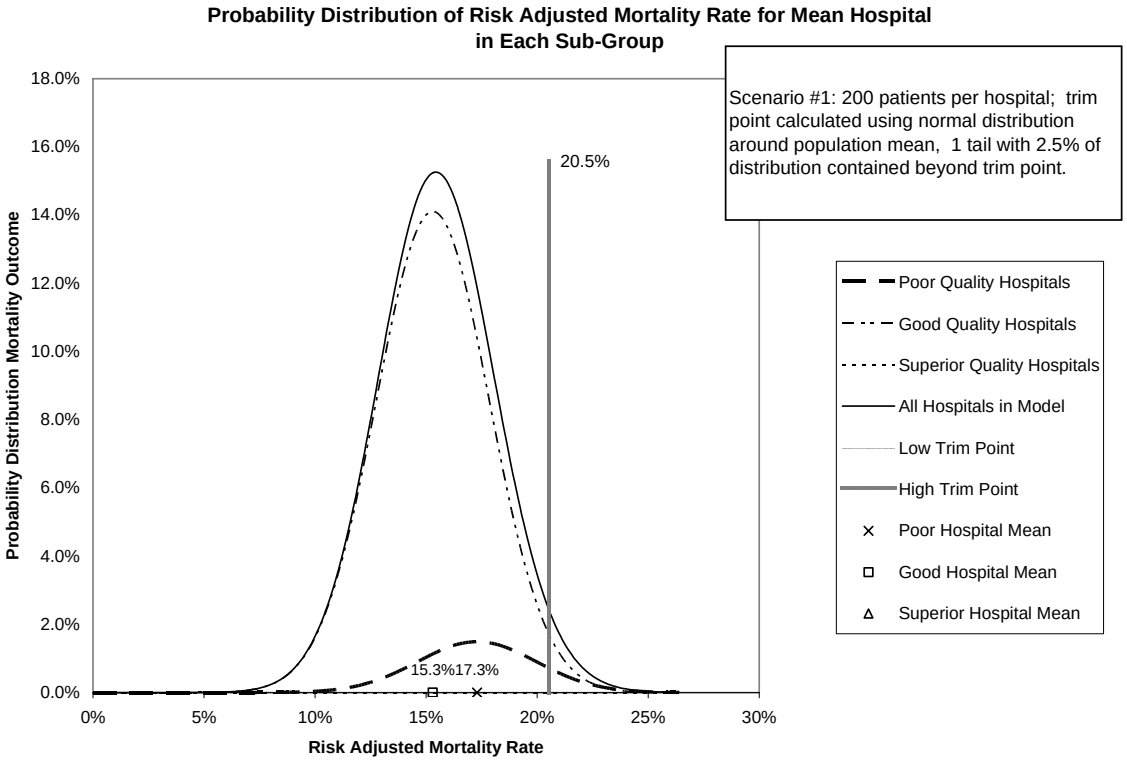
A graphical representation of this hypothetical world of hospitals is shown in Figure 5.

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									

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

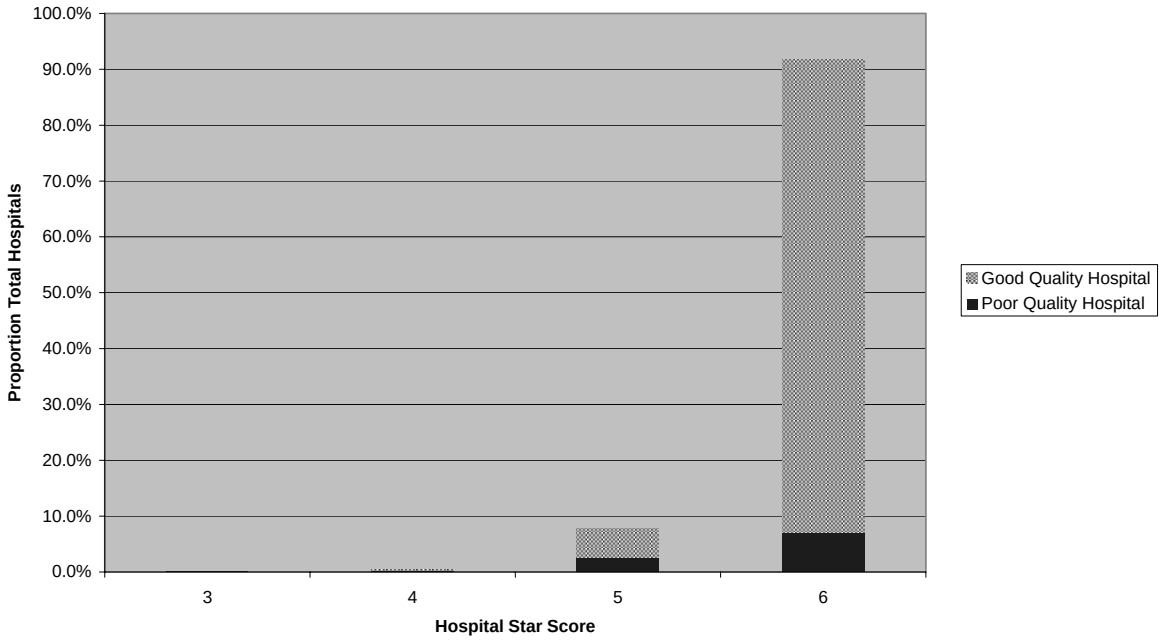


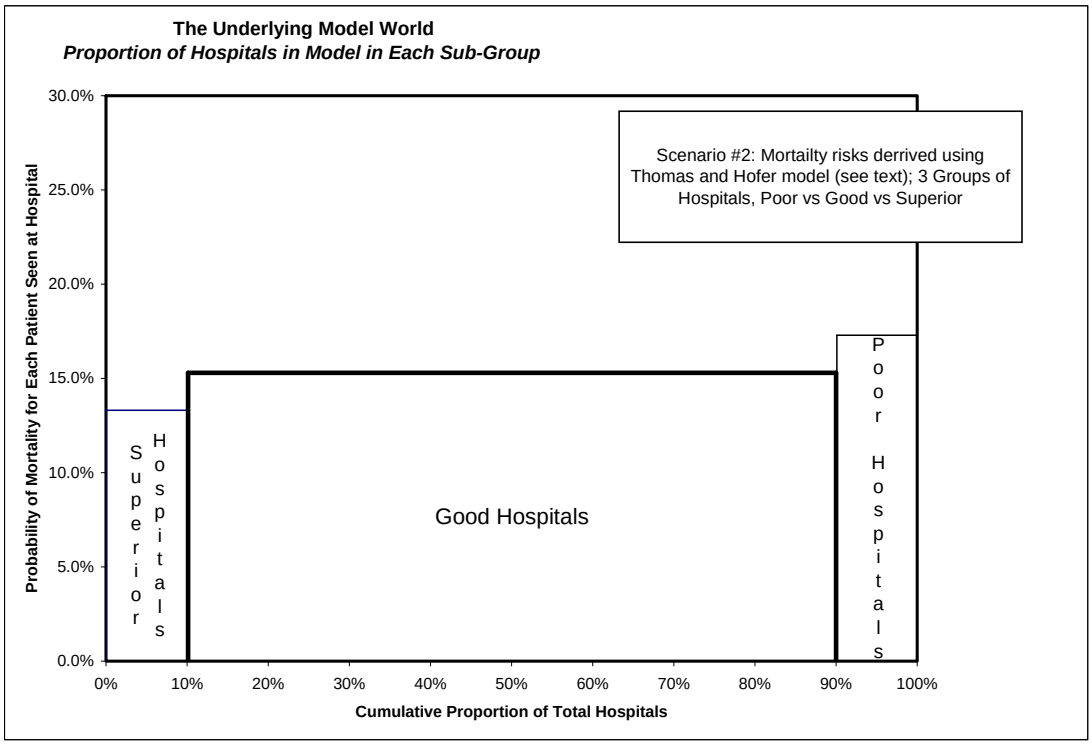
Table 18: Scenario 1: Expected score distribution for good vs. poor hospitals over 3 years

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.

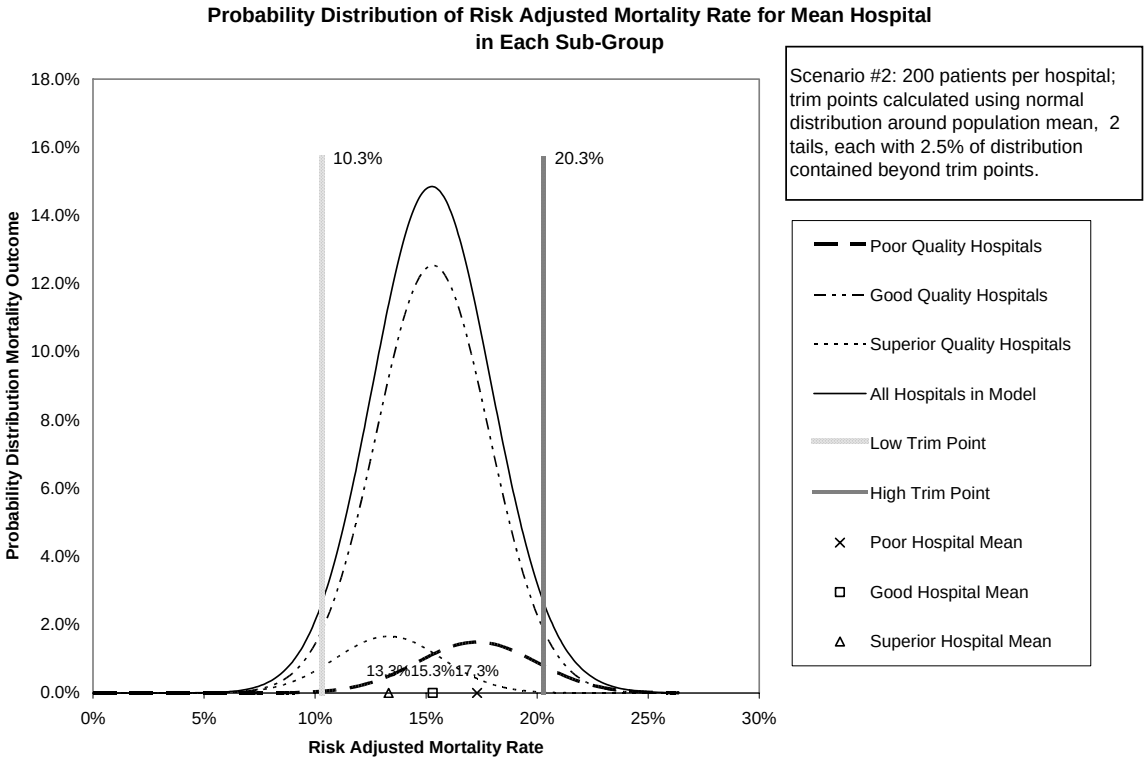
Figure 8: Scenario 2: Hypothetical world of hospitals



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.

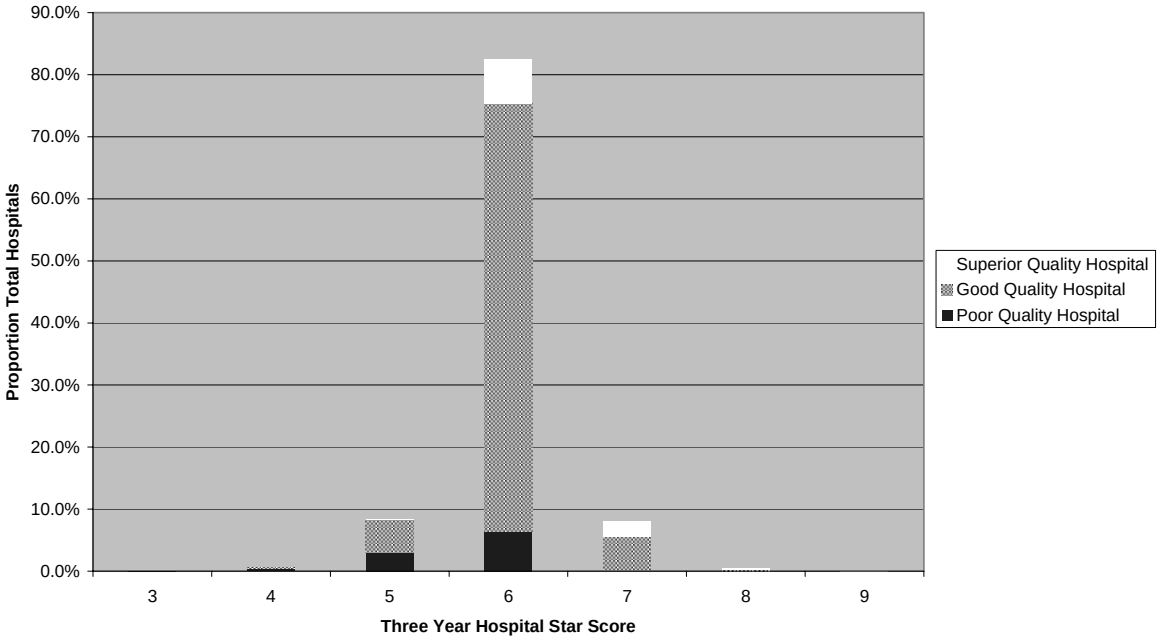
Figure 9: Scenario 2: Hypothetical world and evaluation function



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

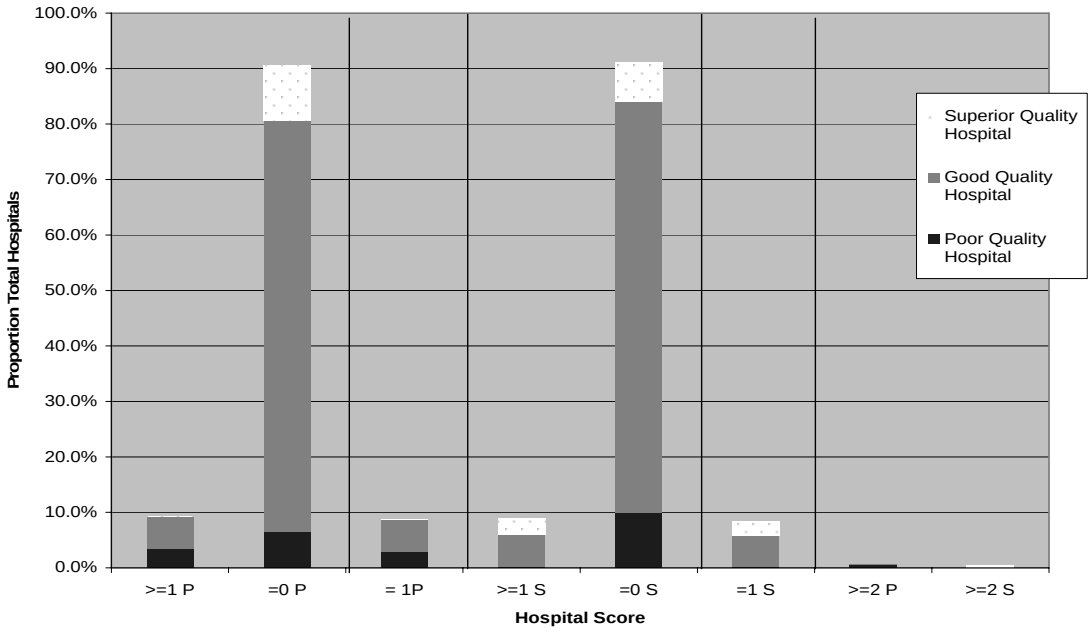


Derivative scores were used to assess whether further discrimination could be obtained among the three sub-groups. The measures are *never poor* (= 0 P), *ever poor* (>= 1 P), *exactly 1 poor* (= 1 P), *mostly poor* (>= 2 P), *never superior* (= 0 S), *ever superior* (>= 1 S), *exactly 1 superior* (=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

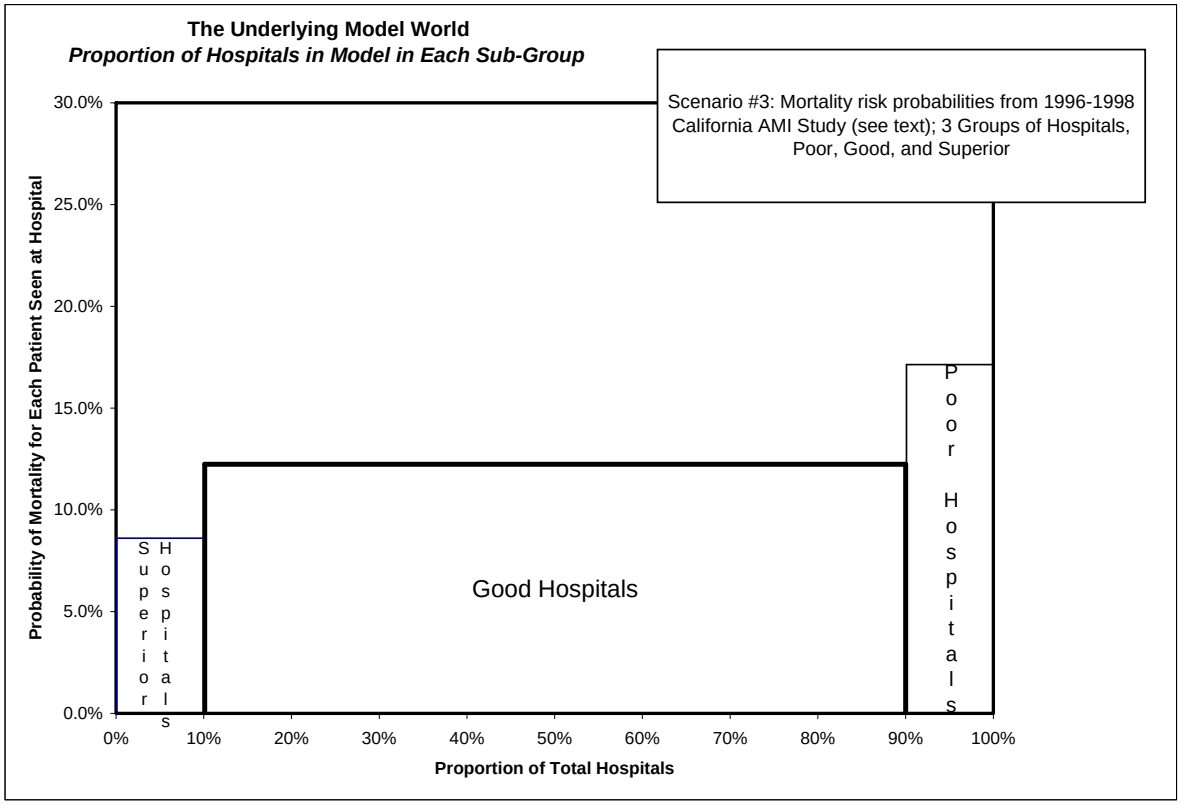
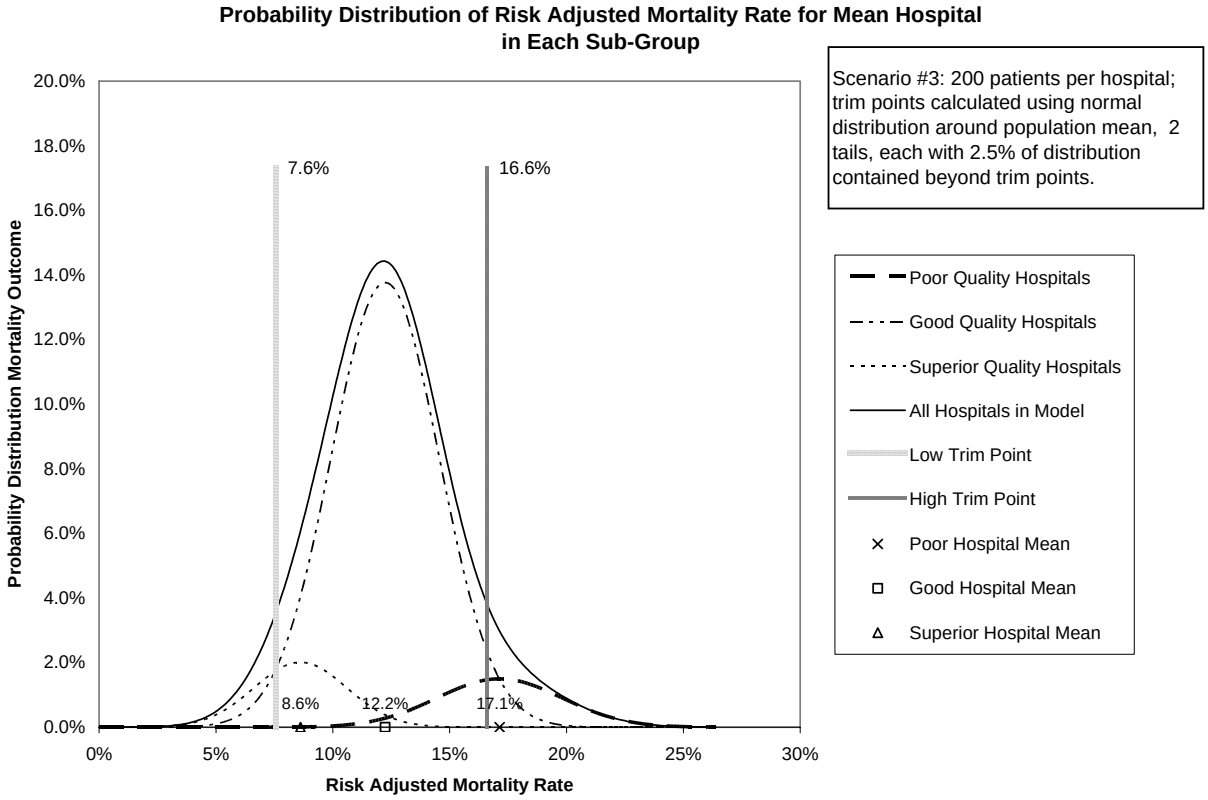
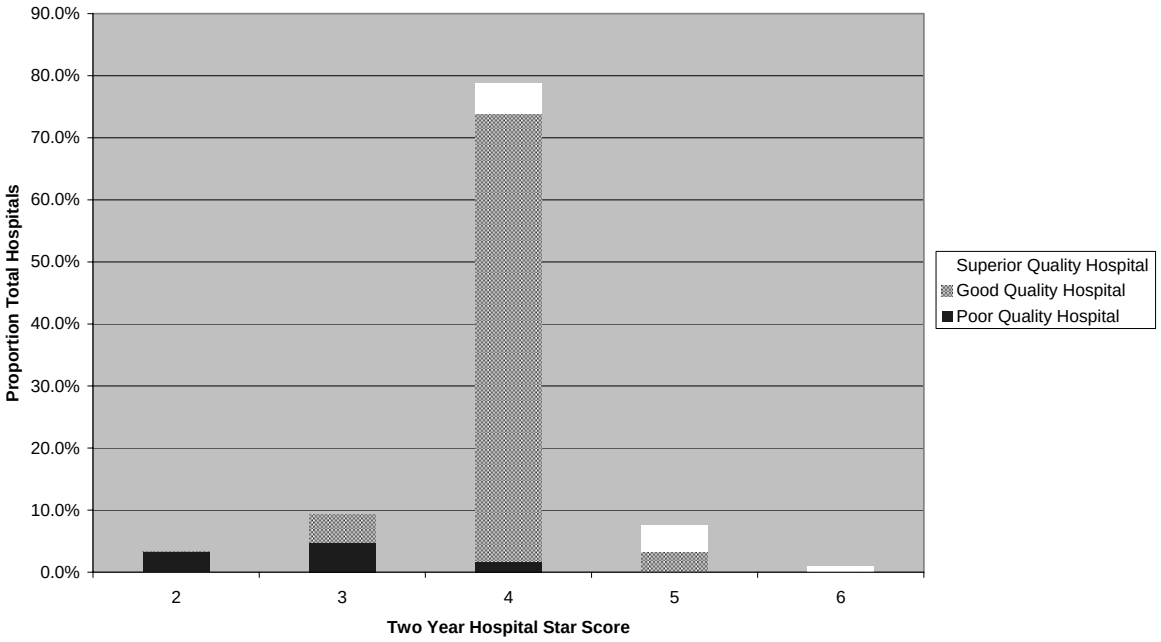


Figure 13: Scenario 3: Hypothetical world and evaluation function



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

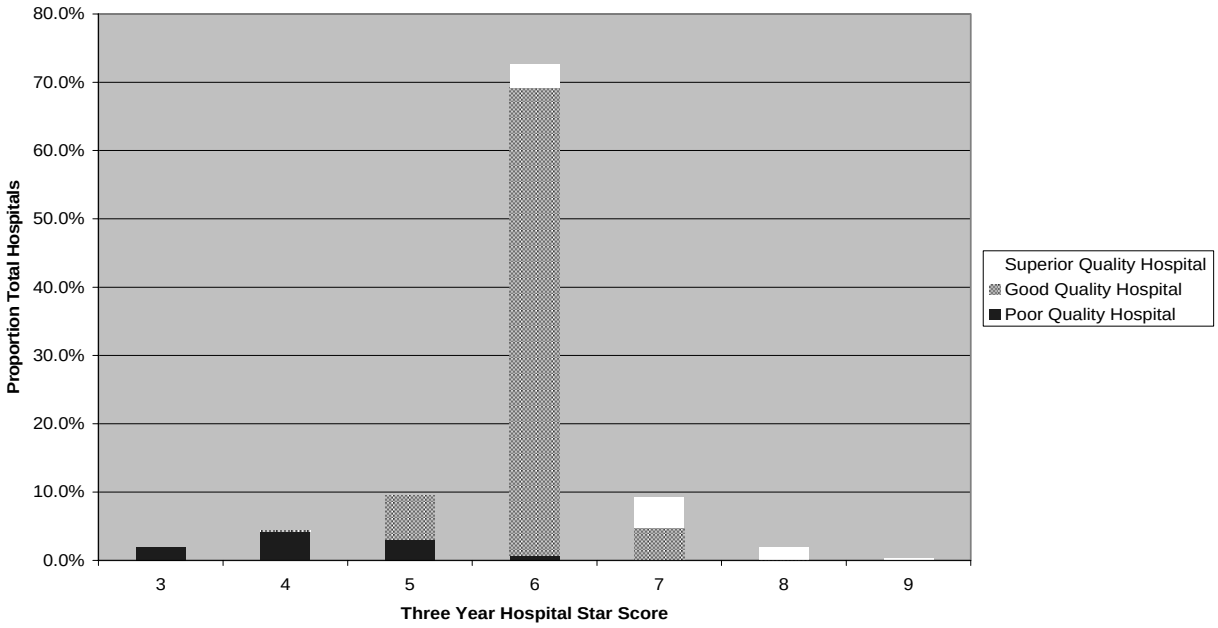


Figure 16: Scenario 3: Three-year derivative scores, predictive values

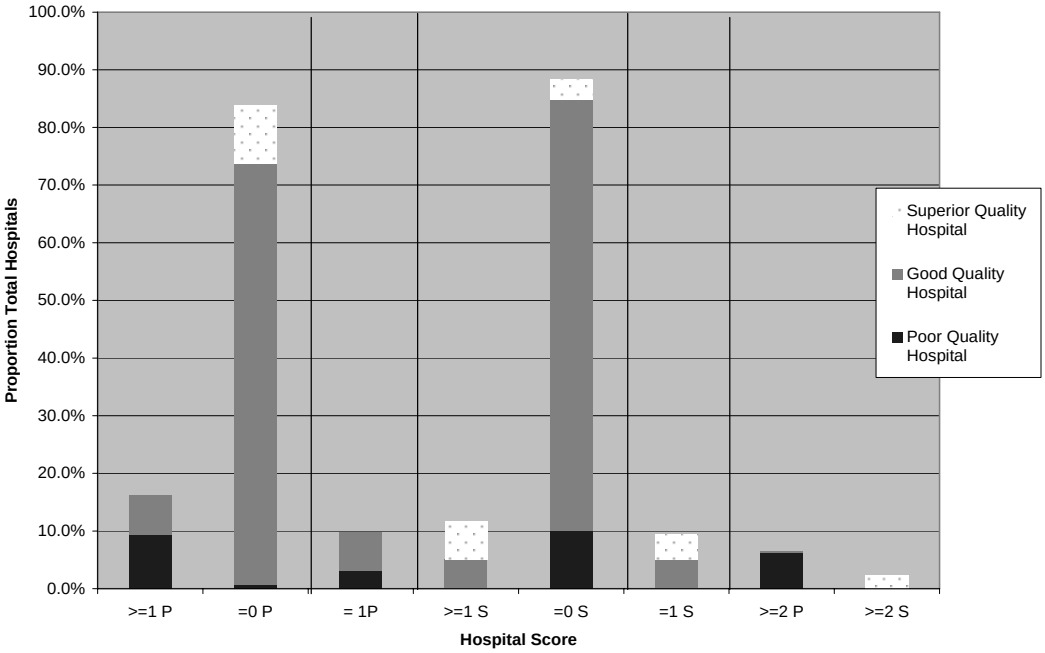
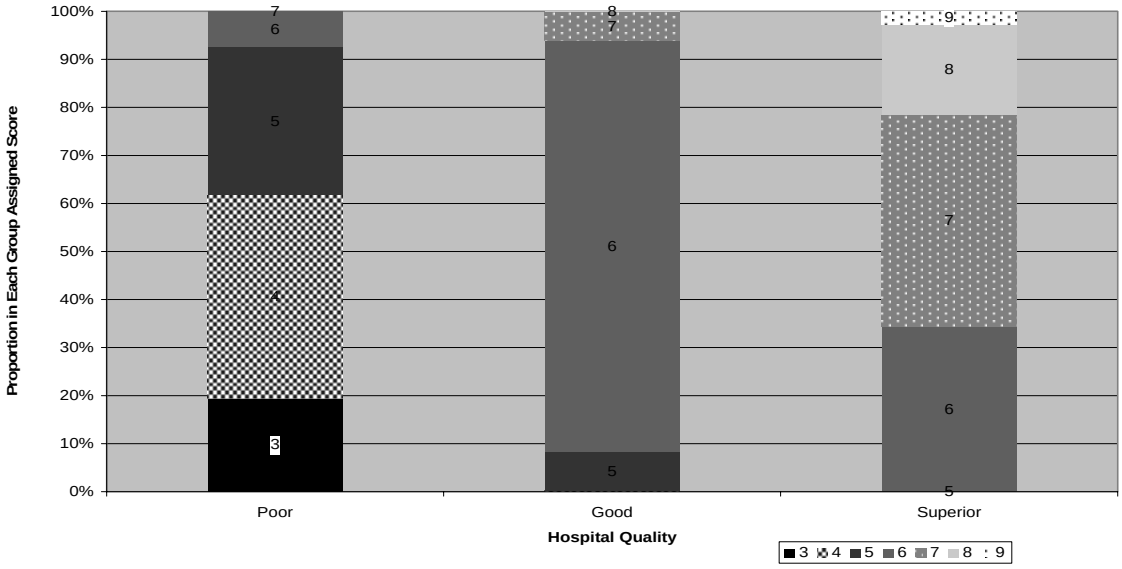


Figure 17: Scenario 3: Distribution of 3-year derivative scores, predictive values



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

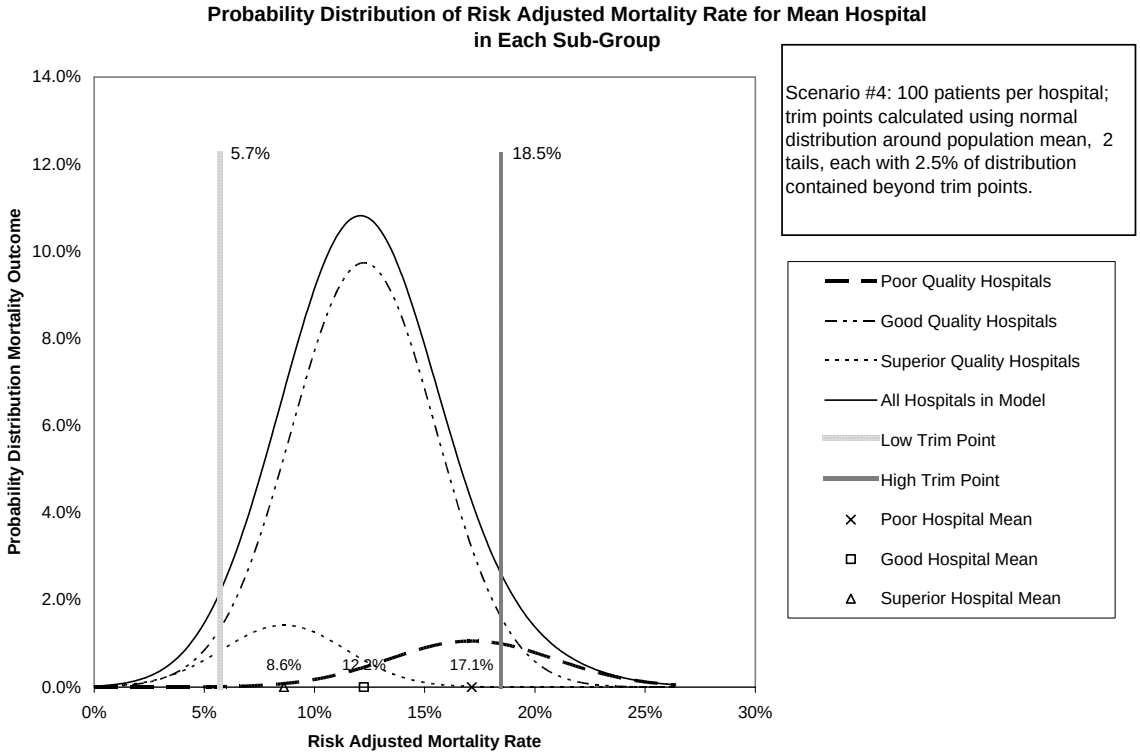
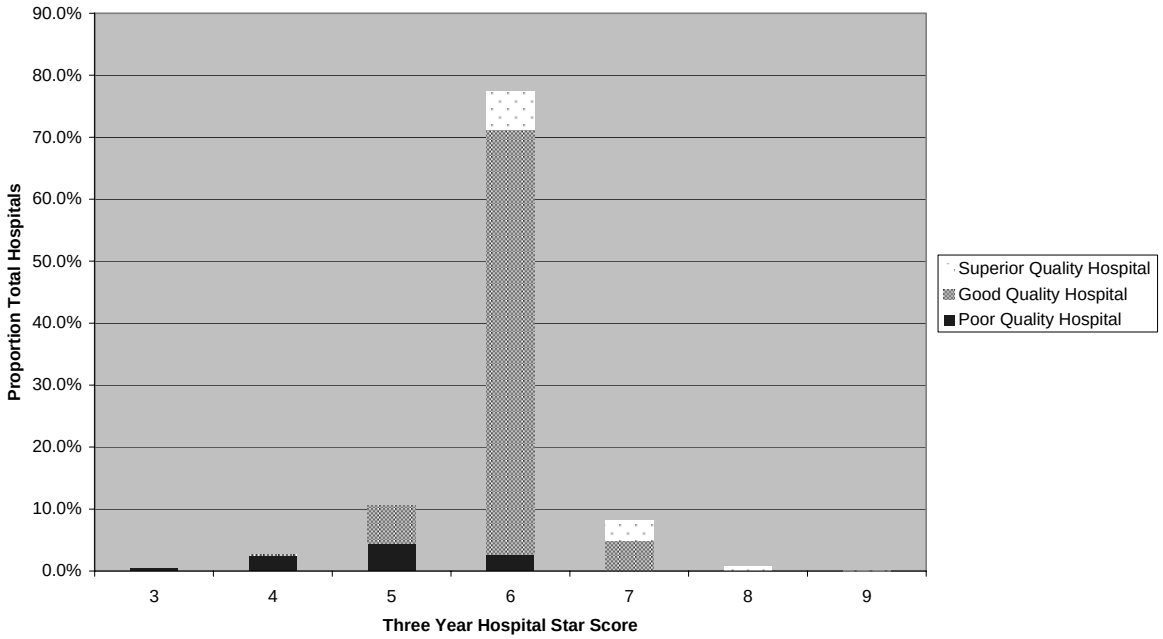


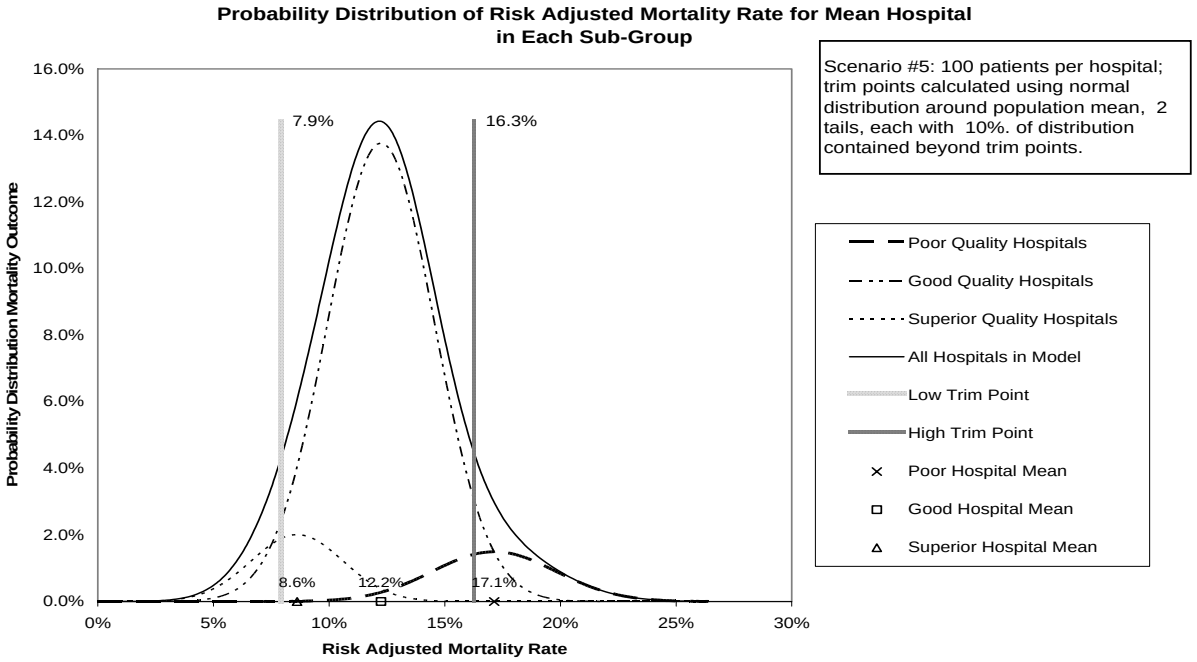
Figure 19: Scenario 4, year 3: Proportion of superior, good, and poor hospitals by 3-year star score



Scenario 5: Identifying a Higher Proportion of Outliers

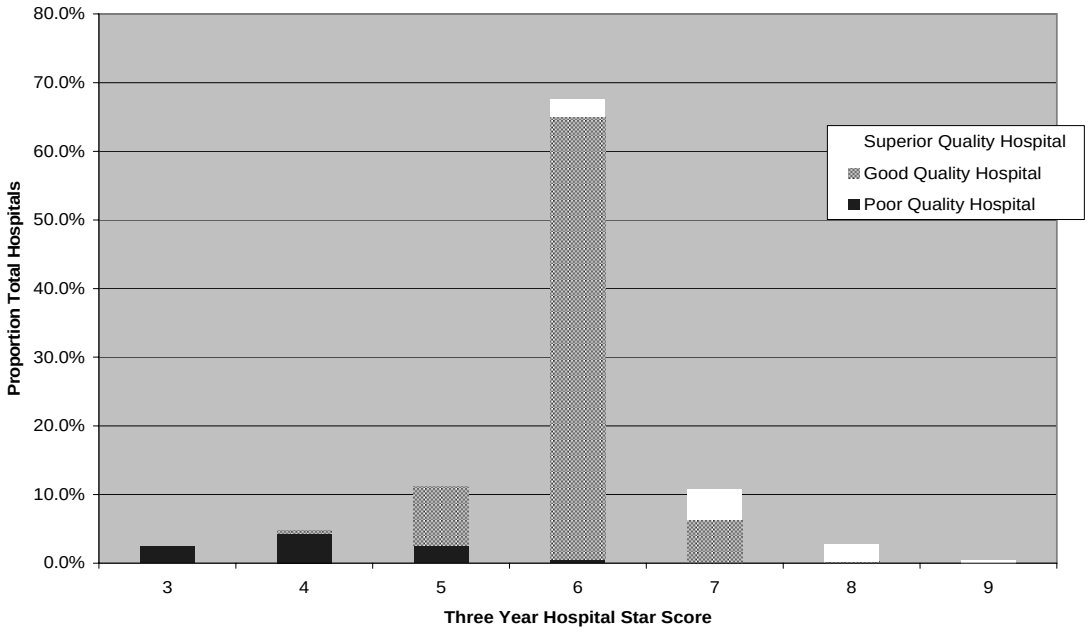
In this simulation, the same hypothetical world as in scenario 3 was used, however, the definition of the trim points for the grading function was changed. In this scenario, the trim points are set such that 10% of the overall hospital quality distribution lies to the right of the upper trim point, and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

6. Discussion

Analysis of Published and Ongoing Research

Performance-based payment and reputational incentives. The available literature about QBP is sparse and there is little evidence base from which to answer the key questions listed in Chapter 2. For those studies that are available, the results are mixed. The incentive strategies used and dependent variables measured are too different among the studies to permit formal meta-analysis. Key variables frequently go unreported, making it more difficult to reach firm conclusions about the potential for and limitations of QBP. Furthermore, since several important variables are not included in any study, the potential for these factors to influence the observed results of these studies is unknown. This means users of the available studies of QBP must be cautious and rely on their judgment in drawing lessons from this literature.

With those caveats, it does appear that in some circumstances, providers respond appropriately to financial incentives. For instance, Hickson et al. show that a financial incentive as small as $2 per visit is enough to increase pediatrics residents' willingness to do well-child care and provide continuity of care.[43] Similarly, Hibbard et al. show that reputational incentives increase the quality improvement activities of hospitals, especially those that are performing poorly.[61]

The optimal approaches to QBP—and the determinants of when one approach is more effective than another—remain uncertain, given the literature. However, some factors identified in the conceptual model do seem to matter. In particular, the observation that significant responses were more likely for quality indicators that reflected clinician performance, rather than patient compliance (e.g., for tobacco screening vs. tobacco cessation) suggest that enabling or inhibiting factors are important. By extension, this implies that the difficulty and cost of achieving the performance goals (both of which rise when patient barriers increase) may also be important determinants of the response to an incentive.

In addition, both studies in which the performance threshold for receiving bonuses was uncertain (because it depended on the performance of other medical groups) were negative. This suggests that uncertainty about revenue potential may be a factor, but strong conclusions cannot be reached based on two studies, even though they are randomized and controlled.

Other potentially important factors have not been studied. These include potentially predisposing factors such as the presence and impact of other community initiatives or incentive programs (which could create an "incentive cacophony"), provider characteristics, and enabling (or inhibiting) factors at the organizational level.

The absence of studies of organizational factors may be particularly important, since responses to some incentives may be determined at the organizational level. For instance, many observers are advocating the use of clinical teams and information systems, both of which would be difficult and expensive for a single provider but might be more feasible for a group. Furthermore, an increasing number of providers are practicing in group settings,[4] so a rising proportion of the priority-setting for, and systems investments by, providers comes from the organizational level**.** The optimal approach to QBP may include a mix of incentives directed at both organizational and individual provider levels, but this has never been studied.

The focus on preventive measures in the available literature likely reflects that these have traditionally been disproportionately represented in accreditation and other data collection

processes (such as HEDIS®) and that they can be measured from administrative data, which is part of the reason they are in HEDIS® and also makes research easier to perform. However, this focus also means that most of the available literature addresses quality problems in the underuse category, rather than overuse or misuse, even though problems of this type are quite common.

Ongoing research into QBP. The research projects that are currently in progress will provide some important information. They should provide an estimate, at least at a point in time, of the extent of use of QBP, and describing several specific projects and the determinants of participation among providers. In addition, several evaluations using contemporaneous (though not randomized) control groups or natural experiments are planned, and these should provide some information about the impact of various QBP strategies. However, the lack of randomized controlled trials makes it extremely likely that there will be continuing questions about each of the QBP strategies tested, especially about whether uncontrolled factors that differ among the intervention and control groups explain the observed results. Moreover, simple trials are not designed to test the effects of a QBP intervention with sufficient sample size to assess whether performance differs across the various predisposing and enabling factors that might affect variations in performance. In essence, they would be analogous to testing whether chemotherapy "works" against cancer, without specifying the nature of the drugs, their regimens, or even the type and stage of the cancers.

Evaluating Outcomes Reports

Our simulations suggest that outcomes reports can yield useful evaluations of hospital and other provider performance. We reach different conclusions from prior investigators for three reasons. First, we assume that, while mislabeling may occur in a single period, it is unlikely to have significant impact on a hospital unless it is repeated over multiple years, which we show would be a very rare event. Second, we introduce the notion of several categories of providers. We believe that it is less important to mislabel a provider from its own category to the adjacent one than it is to miss by multiple categories, and we find that these major mistakes are rare, even with relatively small sample sizes. Finally, by using recent data reflecting the much larger than previously expected differences in outcomes (which have been validated by studies of processes), we have modeled hospital populations with larger differences in underlying mortality rates. These results are consistent with the notion that chance can have an impact on providers' reputations in the short term, but that it should not be a major barrier to outcome reporting if one assumes long term relationships between providers, their patients, local purchasers, and other stakeholders.

In addition, our results show that, despite the statistical "noise" created by random variation, evaluation and labeling systems can be developed that can discriminate poor quality hospitals from good or superior hospitals. Such evaluations, by their nature, will have better grading accuracy when the distributions of the underlying hospitals to be graded (that is, the groups of the hypothetical world) have little overlap. Overlap is reduced when scores are based on outcomes in which the difference between good and poor (or superior and good) performance is large or when the number of patients per hospital is large (to minimize variation due to chance). In cases in which the outcomes in question have overlapping distributions in the hypothetical world, the evaluation system can be improved by using multi-category evaluations (i.e. more than just the labels "good" and "poor") and summary grades over time. Each of these approaches has pros and cons.

Evaluation using multiple categories has the advantage that one can be more assured of accuracy of the grades that differ most from the mean. In our examples, it would be unlikely for a hospital with a 3-year *star* score of 3 (tentatively rated as *poor* 3 consecutive years) to actually be a superior quality hospital. However, multi-category grading does increase the chance of minor mislabeling. It would be a fairly common event for a hospital to receive, for example, *5 stars* in a given 3-year period, even when its long term performance was actually at the *6-star* level. In addition, how the multi-level category scores would be perceived by and interpreted by the users of hospital performance reports is an issue that requires careful thought. In our hypothetical hospital domains, there would not be a reliable difference between hospitals receiving a score of 5 or 6. Yet some stakeholders may tend to order hospitals with these middle scores, despite the lack of reliable differences among them. This is not an uncommon situation in other scoring systems—e.g., *Consumer Reports* frequently indicates that certain products are of approximately the same quality and are listed alphabetically.

Multi-category scores are perhaps most useful in that they can identify a subset of hospitals that are almost definitely truly *superior* or truly *poor* quality. With the former group, one can search for process differences that could form the basis of benchmarking or providing lessons for process improvement at other hospitals. Several processes contributing to improved outcomes were identified by analyzing the reasons for outcome variations among hospitals in New York.[70] Conversely, hospitals with definitely *poor* performance can be studied to search for process-level explanations for their sub-par outcomes. Thus, measuring outcome data may help us learn which processes to change and monitor. Furthermore, hospitals should not, and are not likely to, wait until they receive three consecutive *poor* quality assessments before doing something. While a single *poor* score may just be chance, any reasonable quality improvement team would start examining charts and processes after a second *poor* quality score, if only to be able to report back to the CEO on what they found before the next quality reports come out.

Future Research

Study design issues. From the literature review, it should be clear that pursuing research without a conceptual/theoretical model leads to incomplete reporting of key variables, and research designs that produce results that are not very useful for policy recommendations. Thus, the first requirement for subsequent research should be a clear delineation of how it fits into an overall scheme for testing conceptual models of QBP. For instance, a common (and valid) target for quality improvement is cancer screening. In the short term, cancer screening can be expected to increase utilization (by both the initial testing and the evaluation of positive tests). In most capitation agreements, these additional services would be covered by the capitation fee. Therefore, theory suggests that any evaluation of a QBP program to increase cancer screening in capitated environments should explicitly consider the magnitude of the costs of screening for, diagnosing, and treating more cancer in comparison to the incentive offered. This is not to suggest that the simple economic incentive within capitation to avoid screening costs leads to bad behavior by providers focused on their capitation balance. However, if the organization's quality improvement committee is trying to decide whether to focus limited resources on responding to an incentive to increase screening for cancer or an equivalent incentive to increase physician counseling regarding smoking cessation, the latter might be chosen, because it is less

burdensome in a variety of ways. Only by considering all these costs and barriers to responding to an incentive can its true impact be understood.

There are also important issues of control groups and analytic plans. Though there clearly need to be more randomized controlled trials of QBP, these are difficult and often expensive to undertake. However, as the number of purchasers and health plans adopting QBP increases, there will be more opportunity to use contemporaneous control groups that, though not randomly selected, could be useful, especially if attempts are made to match them to intervention groups in terms of characteristics identified in our conceptual model. As these study designs are more subject to bias than randomized trials, we believe extensive use of qualitative analytic methods will be valuable in augmenting the quantitative analysis of an incentive's impact with participants' and observers' judgments about barriers to and determinants of responses to the incentives.

Topics of investigation. Theory also should play a greater role in the selection of topics to be studied. Since most of the existing research focuses on incentives to individual providers, but the conceptual model suggests that organizations could have a profound influence on performance, a topic needing further investigation is the relative importance of individual versus organizational incentives. In addition, the model suggests the need to address special situations, such as when market characteristics (e.g., local monopolies) are the dominant feature of purchaser-provider relations. This does not imply that all studies must begin with theory—we recognize that in many instances researchers will have to work with the interventions that are being put into place by purchasers. Theory, however, may help inform the selection of intervention goals, of the timing of site involvement, and of the selection of "control" or comparison groups. The theoretical framework we have outlined may also help design better interventions simply by causing people to think more carefully about the incentives, enabling factors, and potential barriers.

Finally, we found only one trial that compared two different QBP approaches; all other studies had a "placebo" control group. A major goal should be to address this weakness with studies that compare performance-based payment to reputational strategies and compare different strategies within the payment and reputational subcategories to each other. These evaluations should include temporal components as well. For instance, it may be that there is some attenuation of response to reputational strategies over time if they are not subsequently backed up with payment incentives.[5]

Planning research programs. While individual research projects should reflect theory, funders may also wish to consider using theory to drive their approach to developing a portfolio of research. In particular, we suggest two general approaches, which we refer to as *sequential hypothesis testing* of incentive strategies and *parallel hypothesis testing* of enabling and predisposing factors. By sequential hypothesis testing we mean that a research program could proceed in a logical fashion from tests of incentives that have a higher probability of being successful (that is, of stimulating performance improvements) toward those that, *a priori*, would be expected to be less likely to be effective.

For instance, consider the QBP strategies of additional fee-for-service payment versus paying bonuses to providers from a fixed pool based on relative performance. There are features about the bonus pool approach that purchasers might find attractive, such as: 1) the total payout can be set in advance, 2) purchasers can raise or lower this figure periodically and precisely as provider performance and market situations change (e.g., after initial investments to improve performance are paid off, providers may need less of an incentive to continue or to make smaller incremental

gains), and 3) other stakeholders can see exactly how large a commitment purchasers are making to quality. On the other hand, most of the current health care environment is fee-for-service and providers may be more willing to accept, or at least more likely to believe they understand the implications of, a fee-for-service approach. They may also be resistant to the program if they feel that even if they improve, their bonus would be jeopardized if someone else improves more—especially if they could argue that the data or baseline states were not comparable. Therefore, it may be reasonable, as a supporter of research, to consider fee-for-service projects that seem feasible initially, with the understanding that even if these work, it does not guarantee that other methods with which providers are less familiar will have similar impact. Thus, a strategy of simply finding whatever works in getting providers used to being "measured" and receiving explicit rewards for improved performance may be more important than finding the "best" QBP method, at least initially.

If findings accrue suggesting incentive programs that had a higher *a priori* chance of success are indeed effective, funders could begin to consider projects that at least initially seem less feasible or less likely to succeed. If the results of these subsequent trials are negative, they do not negate the prior results, but help place bounds on which approaches are effective. On the other hand, if the subsequent trials are positive, they suggest a wide variety of incentive strategies may be useful. Alternatively, if the approaches thought to be most effective do not work, then either they were "sub-clinical dose", or the underlying strategy should be re-thought. Similarly, the absolute magnitude of the incentive may be an issue, in which case it is useful to start with high pre-test probability of success (that is, with fairly large incentives) and move progressively lower to understand what magnitude of incentive is needed to change behavior. In this manner, the field could move sequentially along a spectrum of hypotheses within each conceptual domain of incentive characteristics, delimiting the range along which QBP strategies can succeed.

Understanding the key aspects of alternative incentive approaches will be important, but will take some time. Therefore, we also recommend simultaneous assessment of the impact of the other elements of the conceptual model, predisposing and enabling factors that mediate the response to incentives. For instance, it is very likely that predisposing factors such as the general financial environment and enabling factors at the organizational level will influence performance, regardless of the use (or not) of QBP strategies. To enhance our understanding of both the potential of QBP and the settings in which it is effective, funders might consider supporting parallel programs addressing these other elements of the conceptual model. That is, getting organizations to install improved information systems, or revising the economic incentives against the coordination of care and preventive services, may by themselves be sufficient to lead to improved performance, without any specific QBP incentives.

It may also be useful to consider the results of this parallel research into predisposing and enabling factors when evaluating subsequent QBP proposals. For instance, if research showed that organizations with disease registries had consistently superior performance, funders might consider whether subsequent QBP trials should be limited to organizations that have registries for the conditions for which performance is measured. It is also important to recognize that certain features may be crucial for some interventions and not others. Registries may be critical to assure that appropriate care is given to patients with diabetes or hypertension, because insufficient contact with the medical care system may be especially problematic for these patients; registries are unlikely to be needed to assure that beta blockers and aspirin are appropriately recommended upon discharge after a myocardial infarction. This combination of

sequential hypothesis testing of incentive strategies with parallel hypothesis testing of other elements of the conceptual model is likely to advance the field much more rapidly than has occurred to date.

The basic tools of performance measurement. Another barrier to QBP is that the science of performance measurement is still underdeveloped.[5, 71] The available set of metrics is not broadly representative of all care, while purchasers must pay for care across the entire clinical spectrum. Furthermore, there has been little research addressing non-clinical outcomes such as absenteeism that may be very important to employer purchasers.[71] Experience in other industries has shown that developing performance measures for complex phenomena is difficult and that inappropriate measures can have significant negative consequences.[72] This suggests that research into QBP should be accompanied by further development of the basic tools of performance measurement.

Conclusion

The environment in which purchasers and providers interact is rapidly changing. There is clearly growing interest in QBP and some evidence that both payment and reputational incentives can work, but, to date, there is little unequivocal data on which to base QBP strategy selection. Fortunately, our modeling suggests that, with appropriate caution, outcomes measures can be included among the performance indicators used for QBP. Furthermore, the notion of using incentives to encourage high quality (as well as actually measuring quality) is much more acceptable than it was a few years ago, and this has increased the number of opportunities to study QBP. Researchers have responded with a broad portfolio of ongoing research that promises to both outline current trends in the use of QBP and offer some preliminary evaluations of several different incentive approaches. Policymakers should expect additional research, especially if designed and selected for funding based on conceptual considerations such as those we outline, to rapidly advance our understanding of how to use performance measurement and incentives to improve the quality of health care Americans receive.

References and Included Studies

1. Kohn LT, Corrigan J, Donaldson M, (eds). *To Err is Human: Building a Safer Health Care System.* Washington, D.C: National Academy Press; 1999.

2. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. *N Engl J Med.* 2003;348(22):2218-2227.

3. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. *N Engl J Med.* 2003;348(26):2635-2645.

4. Dudley RA, Luft HS. Managed care in transition. *N Engl J Med.* 2001;344(14):1087-1092.

5. Mehrotra A, Bodenheimer T, Dudley RA. Employers' efforts to measure and improve hospital quality: determinants of success. *Health Affairs.* 2003;22(2):60-71.

6. Bindman AB. Can physician profiles be trusted? *JAMA.* 1999;281(22):2142-2143.

7. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "Report Cards" for assessing the costs and quality of care of a chronic disease. *JAMA.* 1999;281:2098-2105.

8. 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.

9. Marshall MN, Shekelle PG, Leatherman S, Brook RH. The public release of performance data: what do we expect to gain? A review of the evidence. *JAMA.* 2000;283(14):1866-1874.

10. Marshall MN, Shekelle PG, Davies HT, Smith PC. Public reporting on quality in the United States and the United Kingdom. *Health Affairs.* 2003;22(3):134-148.

11. McLaughlin C. Health Care Consumers: Choices and Constraints. *Med Care Res Rev.* 1999;56(Suppl 1):24-59.

12. Galvin R, Milstein A. Large employers' new strategies in health care. *N Engl J Med.* 2002;347(12):939-942.

13. Wetzel S, Galvin R, Buck CJ, et al. Taking a giant leap forward in promoting quality. *Health Aff.* 2000;19(2):275-276.

14. Jencks SF, Huff ED, Cuerdon T. Change in the quality of care delivered to medicare beneficiaries, 1998-1999 to 2000-2001. *JAMA.* 2003;289:305-312.

15. Jack W. Contracting for health services: an evaluation of recent reforms in Nicaragua. *Health Policy and Planning.* 2003;18(2):195-204.

16. Eichler R, Auxilia P, Pollock J. Performance-based payment to improve the impact of health services: evidence from Haiti. *World Bank Institute Online Journal*. April 15, 2001. Available at: www.worldbank.org/wbi/healthflagship/oj_haiti.pdf. Accessed February 23, 2004.

17. Corrigan JM, Donaldson MS, Kohn LT, (eds). *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, D.C: National Academy Press; 2001.

18. Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. *Am J Public Health.* Nov 1998;88(11):1699-1701.

19. Miller R, Luft H. HMO plan performance update: an analysis of the literature, 1997-2001. *Health Aff.* 2002;21(4):63-86.

20. Dudley RA, Miller RH, Korenbrot TY, Luft HS. The impact of financial incentives on quality of health care. *Milbank Q.* 1998;76(4):649-686, 511.

21. Dranove D, White WD. Agency and the organization of health care delivery. *Inquiry.* 1987;24(4):405-415.

22. Robinson JC. Theory and practice in the design of physician payment incentives. *Milbank Q.* 2001;79(2):149-177.

23. Hellinger FJ. The effect of managed care on quality: a review of recent evidence. *Arch Intern Med.* Apr 27 1998;158(8):833-841.

24. Hutchison B, Birch S, Hurley J, Lomas J, Stratford-Devai F. Do physician-payment mechanisms affect hospital utilization? A study of Health Service Organizations in Ontario. *CMAJ.* 1996;154(5):653-661.

25. Andersen RM. Revisiting the behavioral model and access to medical care: Does it matter? *Journal of Health and Social Behavior.* 1995;36(1):1-10.

26. Town R, Kralewski J, Wholey DR, Dowd B. Assessing the influence of incentives on physicians and medical groups. issue of Med Care Res Rev. 2004 (September).

27. Eisenberg JM. *Doctors' decisions and the cost of medical care: the reasons for doctors' practice patterns and ways to change them*. Ann Arbor, MI: Health Administration Press; 1986.

28. Feinstein AR. The 'chagrin factor' and qualitative decision analysis. 145 (7):1257-9. *Archives of Internal Medicine.* 1985;145(7):1257-1259.

29. Loomes G, Sugden R. Regret theory: An alternative theory of rational choice under uncertainty. 92 (368):805-824. *Economic Journal.* 1982;92(368):805-824.

30. Frey BS. On the Relationship between Intrinsic and Extrinsic Work Motivation. *International Journal of Industrial Organization.* 1997;15:427-439.

31. Kuhn M. *Quality in Primary Care: Economic Approaches to Analysing Quality-Related Physician Behavior*. London: Office of Health Economics; 2003 (August).

32. Christensen DB, Hansen RW. Characteristics of pharmacies and pharmacists associated with the provision of cognitive services in the community setting. *J Am Pharm Assoc.* (Sept-Oct) 1999;39(5):640-649.

33. Christensen DB, Holmes G, Fassett WE, et al. Influence of a financial incentive on cognitive services: CARE project design/implementation. *J Am Pharm Assoc (Wash).* Sep-Oct 1999;39(5):629-639.

34. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. Performance-based physician reimbursement and influenza immunization rates in the elderly. The Primary-Care Physicians of Monroe County. *Am J Prev Med.* Feb 1998;14(2):89-95.

35. Amundson G, Solberg LI, Reed M, Martini EM, Carlson R. Paying for quality improvement: compliance with tobacco cessation guidelines. *Jt Comm J Qual Saf.* Feb 2003;29(2):59-65.

36. Newacheck PW, McManus MA, Gephart J. Health insurance coverage of adolescents: a current profile and assessment of trends. *Pediatrics.* Oct 1992;90(4):589-596.

37. Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. *Ambul Pediatr.* Jul-Aug 2001;1(4):206-212.

38. Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. *Pediatrics.* Oct 1999;104(4 Pt 1):931-935.

39. Grumbach K, Osmond D, Vranizan K, Jaffe D, Bindman AB. Primary care physicians' experience of financial incentives in managed-care systems. *N Engl J Med.* Nov 19 1998;339(21):1516-1521.

40. Hadley J, Mitchell JM, Sulmasy DP, Bloche MG. Perceived financial incentives, HMO market penetration, and physicians' practice styles and satisfaction. *Health Serv Res.* Apr 1999;34(1 Pt 2):307-321.

41. Pantilat SZ, Chesney M, Lo B. Effect of incentives on the use of indicated services in managed care. *West J Med.* Mar 1999;170(3):137-142.

42. Hillman AL, Pauly MV, Kerstein JJ. How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations? *N Engl J Med.* Jul 13 1989;321(2):86-92.

43. Hickson GB, Altemeier WA, Perrin JM. Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study. *Pediatrics.* Sep 1987;80(3):344-350.

44. Landon BE, Reschovsky J, Reed M, Blumenthal D. Personal, organizational, and market level influences on physicians' practice patterns: results of a national survey of primary care physicians. *Med Care.* Aug 2001;39(8):889-905.

45. Mort EA, Edwards JN, Emmons DW, Convery K, Blumenthal D. Physician response to patient insurance status in ambulatory care clinical decision-making. Implications for quality of care. *Med Care.* Aug 1996;34(8):783-797.

46. Chaix-Couturier C, Durand-Zaleski I, Jolly D, Durieux P. Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. *Int J Qual Health Care.* Apr 2000;12(2):133-142.

47. Wee CC, Phillips RS, Burstin HR, et al. Influence of financial productivity incentives on the use of preventive care. *Am J Med.* Feb 15 2001;110(3):181-187.

48. Krasnik A, Groenewegen PP, Pedersen PA, et al. Changing remuneration systems: effects on activity in general practice. *BMJ.* 1990;300(6741):1698-1701.

49. Conrad DA, Cave SH, Lucas M, et al. Community care networks: linking vision to outcomes for community health improvement. *Med Care Res Rev.* 2003;60(4 Suppl):95-129.

50. Fisher ES, Wennberg JE. Health care quality, geographic variations, and the challenge of supply-sensitive care. *Perspect Biol Med.* 2003;46(1):69-79.

51. Baker LC. Managed care spillover effects. *Annu Rev Public Health.* 2003;24:435-456.

52. Nyman JA, Akhtar MR, Feldman R. Does publishing the parameters that trigger review of Medicare claims change provider behavior? Results of the parameter release study. *Med Care.* Oct 1995;33(10):1022-1034.

53. Hellinger FJ. The impact of financial incentives on physician behavior in managed care plans: a review of the evidence. *Med Care Res Rev.* Sep 1996;53(3):294-314.

54. Hoopmann M, Schwartz FW, Weber J. Effects of the German 1993 health reform law upon primary care practitioners' individual performance: results from an empirical study in sentinel practices. *J Epidemiol Community Health.* 1995;49(Suppl 1):33-36.

55. Lurie N, Christianson J, Finch M, Moscovice I. The effects of capitation on health and functional status of the Medicaid elderly. A randomized trial. *Ann Intern Med.* Mar 15 1994;120(6):506-511.

56. Born PH, Simon CJ. Patients and profits: the relationship between HMO financial performance and quality of care. *Health Aff (Millwood).* Mar-Apr 2001;20(2):167-174.

57. Stern RS, Juhn PI, Gertler PJ, Epstein AM. A comparison of length of stay and costs for health maintenance organization and fee-for-service patients. *Arch Intern Med.* May 1989;149(5):1185-1188.

58. Walley T, Murphy M, Codd M, Johnston Z, Quirke T. Effects of a monetary incentive on primary care prescribing in Ireland: changes in prescribing patterns in one health board 1990-1995. *Pharmacoepidemiol Drug Saf.* Dec 2000;9(7):591-598.

59. Davidson SM, Manheim LM, Werner SM, Hohlen MM, Yudkowsky BK, Fleming GV. Prepayment with office-based physicians in publicly funded programs: results from the Children's Medicaid Program. *Pediatrics.* Apr 1992;89(4 Pt 2):761-767.

60. Roski J, Jeddeloh R, An L, et al. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. *Prev Med.* Mar 2003;36(3):291-299.

61. Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood).* Mar-Apr 2003;22(2):84-94.

62. Giuffrida A, Gosden T, Forland F, et al. Target payments in primary care: effects on professional practice and health outcome (Cochrane Review). *The Cochrane Library.* 2002(4).

63. Gosden T, Forland F, Kristiansen IS, et al. Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians. *Cochrane Database Syst Rev.* 2000(3):CD002215.

64. Jarlier A, Charvet-Protat S. Can improving quality decrease hospital costs? *Int J Qual Health Care.* Apr 2000;12(2):125-131.

65. Seidman JJ, Bass EP, Rubin HR. Review of studies that compare the quality of cardiovascular care in HMO versus non-HMO settings. *Med Care.* Dec 1998;36(12):1607-1625.

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

67. 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.

68. 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.

69. 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.

70. Dziuban SWJ, McIlduff JB, Miller SJ, Dal Col RH. How a New York cardiac surgery program uses outcomes data. *Ann Thorac Surg.* 1994;58(6):1871-1876.

71. Scanlon DP. Overcoming Barriers to managing Health and Productivity in the Workplace. In: Kessler RC, Stang PD, eds. *Health and Productivity: Emerging Issues in Research & Policy*. Chicago: University of Chicago Press; in press.

72. Ittner CD, Larcker DF. Coming up short on nonfinancial performance measurement. *Harv Bus Rev.* 2003;81(11 (Nov)):88-95, 139.

Acronyms and Abbreviations

AHRQ	Agency for Healthcare Research and Quality
CHCF	California HealthCare Foundation
EMR	electronic medical record
FFS	fee for service
GOLD	Grants On-Line Database
GP	general practitioner
HEDIS®	Health Plan Employer Data and Information Set
HMO	health maintenance organization
IDS	integrated delivery system
IOM	Institute of Medicine
JCAHO	Joint Commission on Accreditation of Healthcare Organizations
MESH	medical subject heading
MI	myocardial infarction
PCP	primary care provider
POS	point of service
PPO	preferred provider organization
QBP	quality-based purchasing
RWJF	Robert Wood Johnson Foundation
VBP	value-based purchasing

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.)

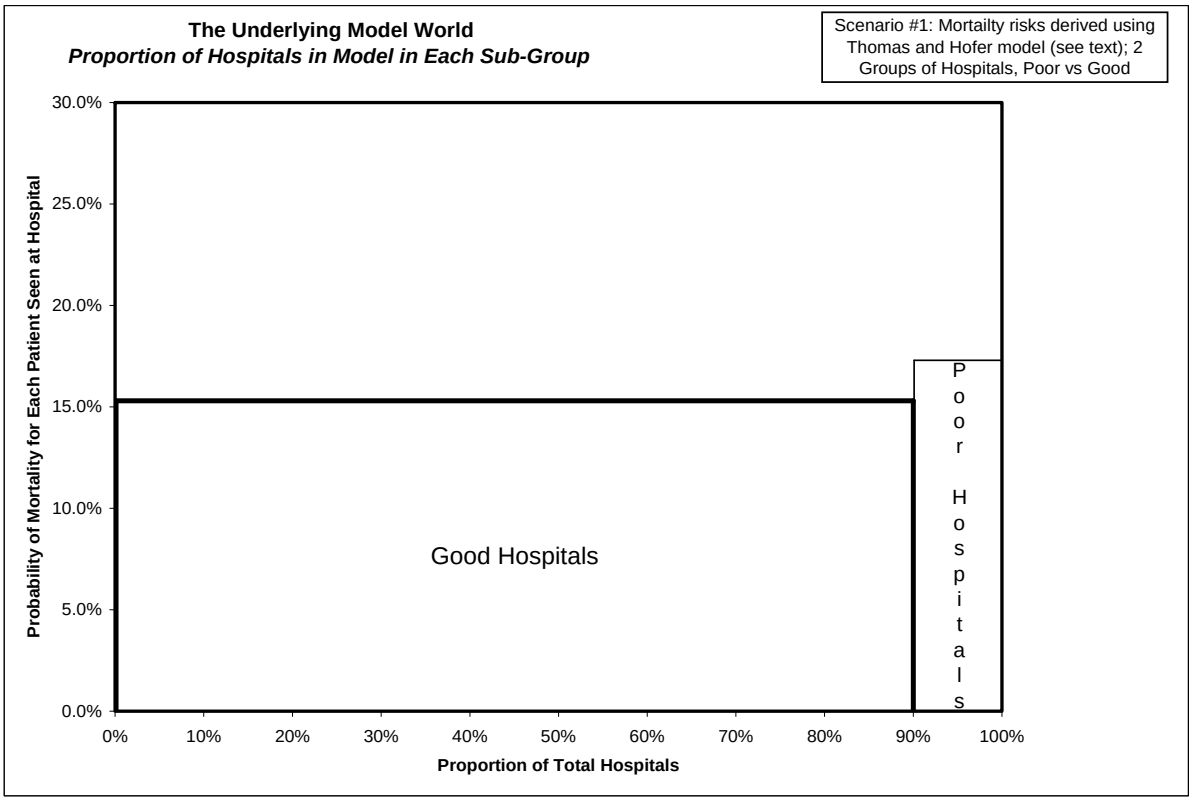
The second step is to calculate, given the first assumptions, the probability that an individual hospital with known characteristics will receive a particular label (e.g., "poor" vs. "good" vs. "superior") and how often those labels will be misapplied (e.g., that a poor quality hospital will be labeled "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

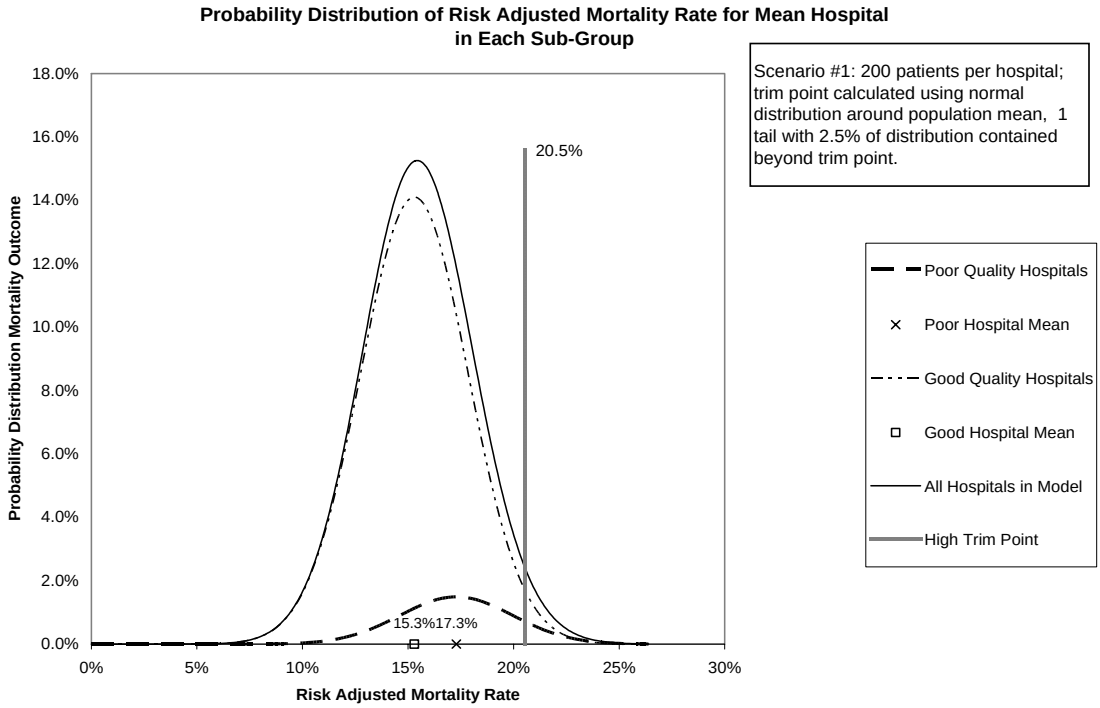


To label hospitals, Thomas and Hofer use an evaluation system similar to clinical diagnostic tests. They define poor performance as that which would be found in the tails of a distribution normally distributed about the mean hospital performance. In their trials, they used a 5% cutoff, so performance 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).

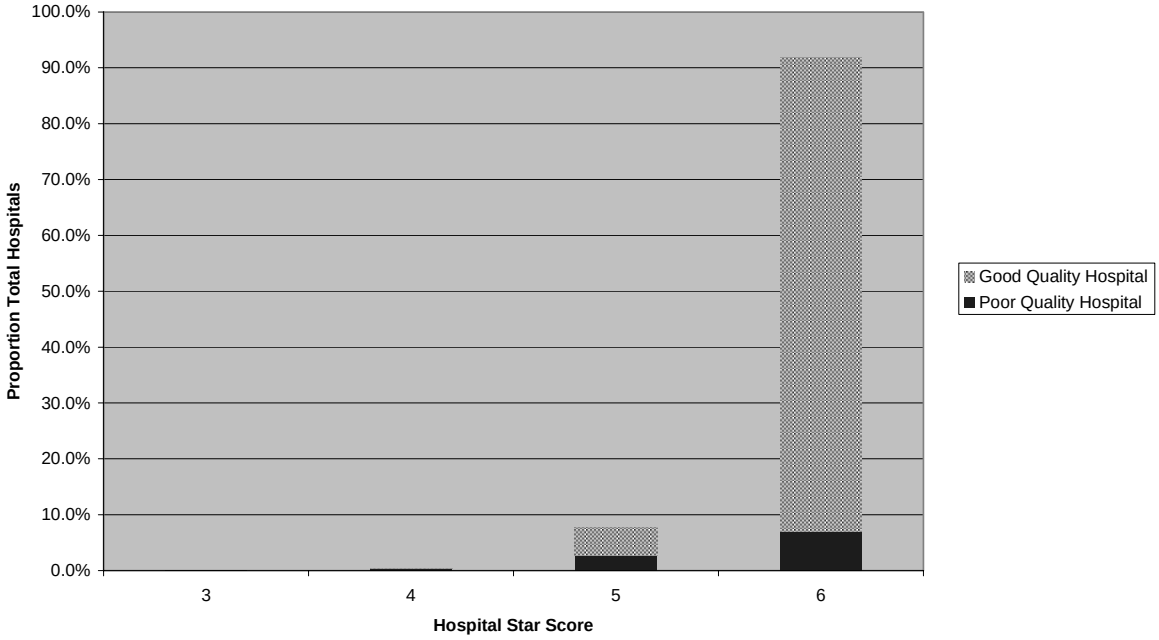
Sensitivity and specificity calculations show that specificity of *4 stars* is 96.1% and sensitivity of *2 stars* is only 1.2%, as 2 stars is very unlikely in this scenario, whether the hospital is poor or good.

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

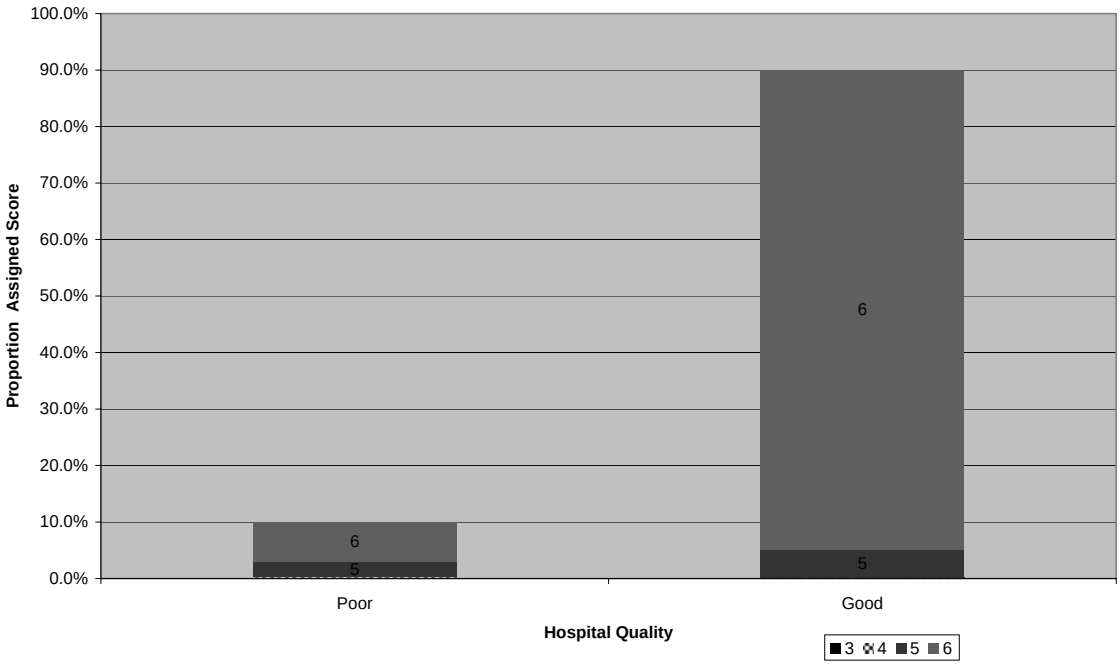


Hospitals with *3* or *4 stars* are almost certainly of *poor* quality—but these scores are rare. Indeed, it is a rare thing to be graded *poor* in this scenario, and to have it occur even once in 3 years happens for only 8.2% of hospitals.

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%

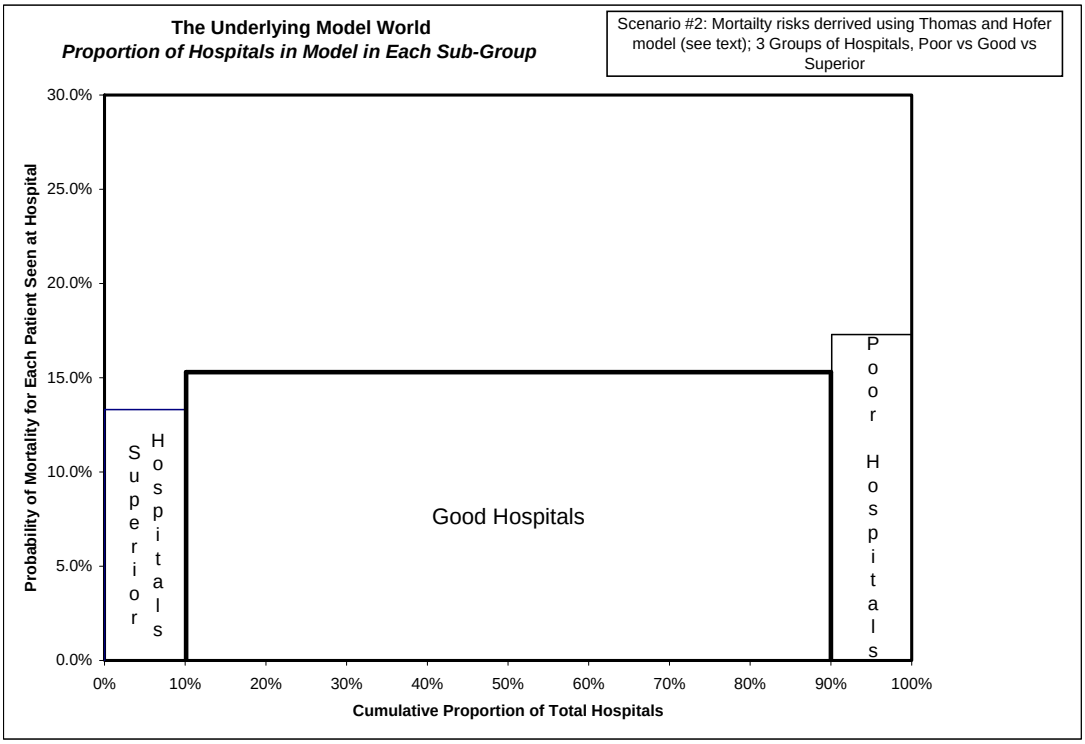
Figure C 4: Scenario 1: Expected 3-Year Score Distribution for Good vs. Poor Hospitals



Scenario 2: Adding Another Hospital Category

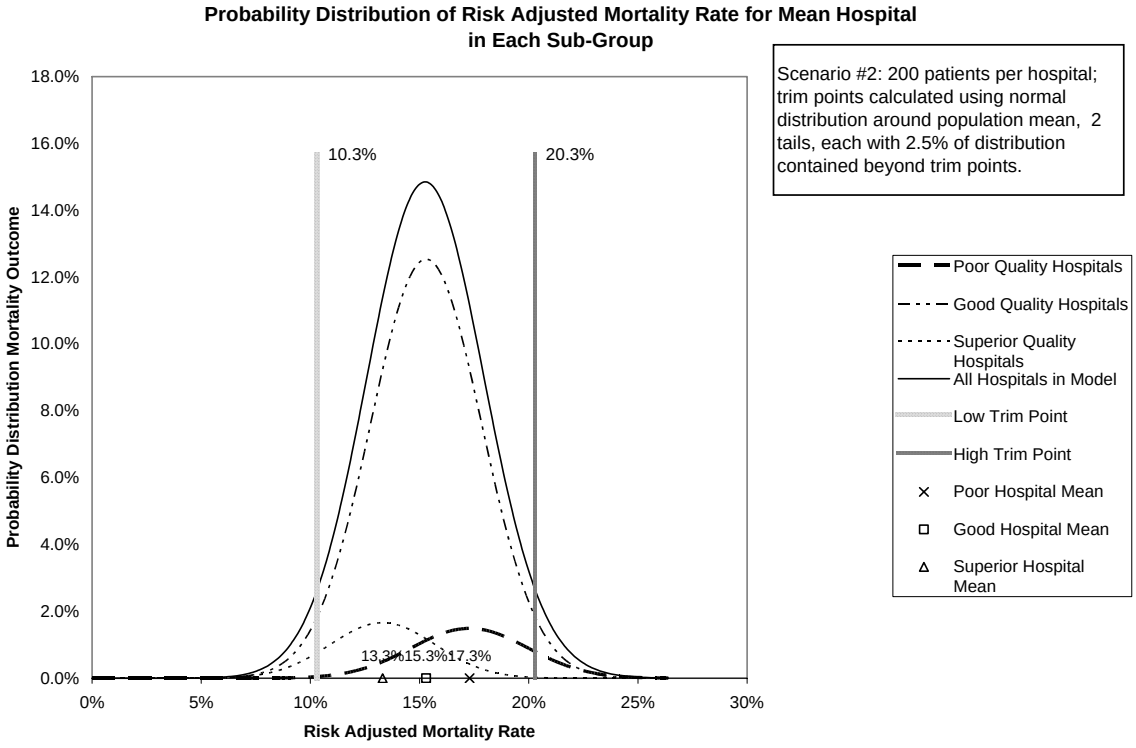
For this scenario, we added the *superior* quality hospital group as 10% of the hypothetical hospital population. The average mortality rate for *superior* hospitals was assumed to be the same percentage difference below the mean performance as Thomas and Hofer's *poor* quality hospitals were above the 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

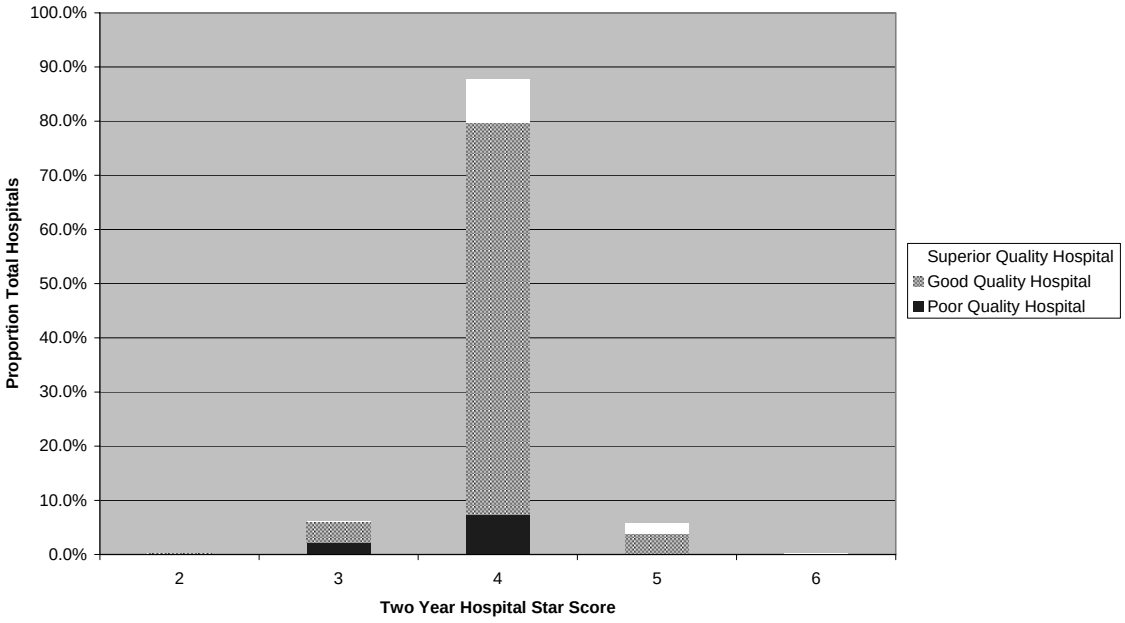
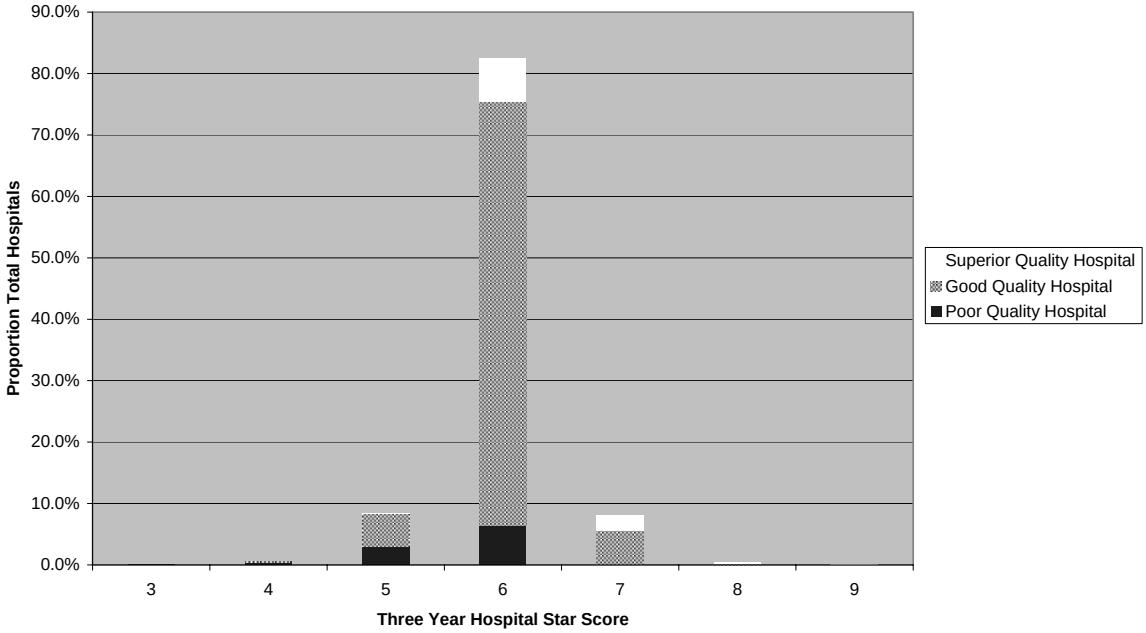
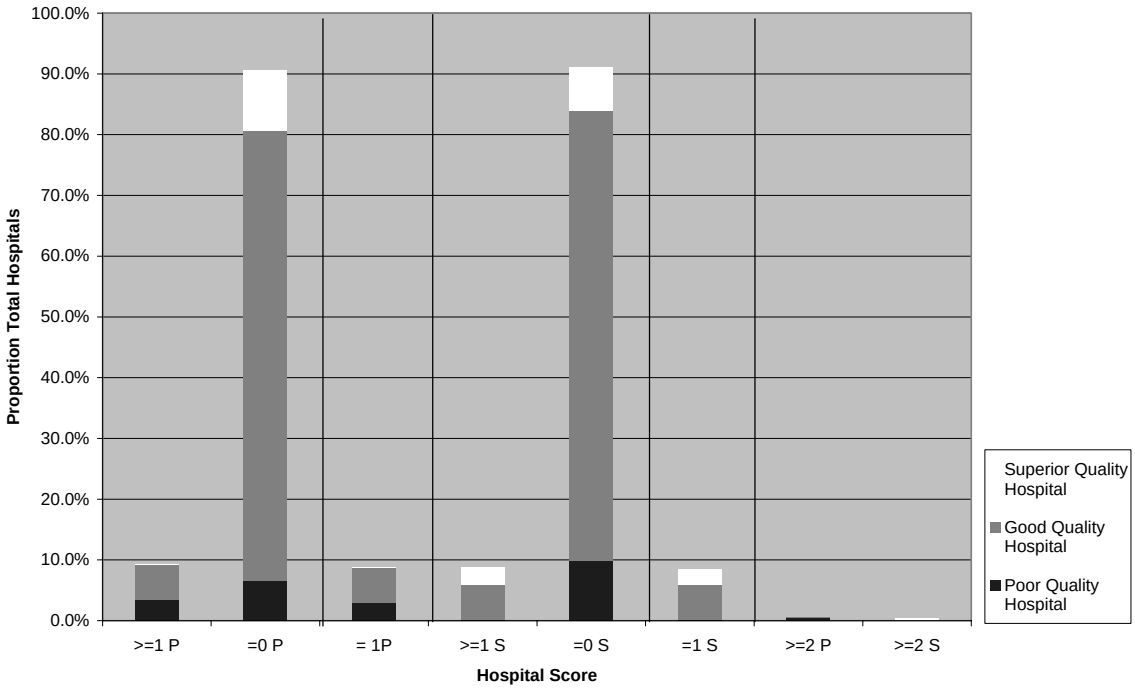


Figure C 8: Scenario 2: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score



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**.**

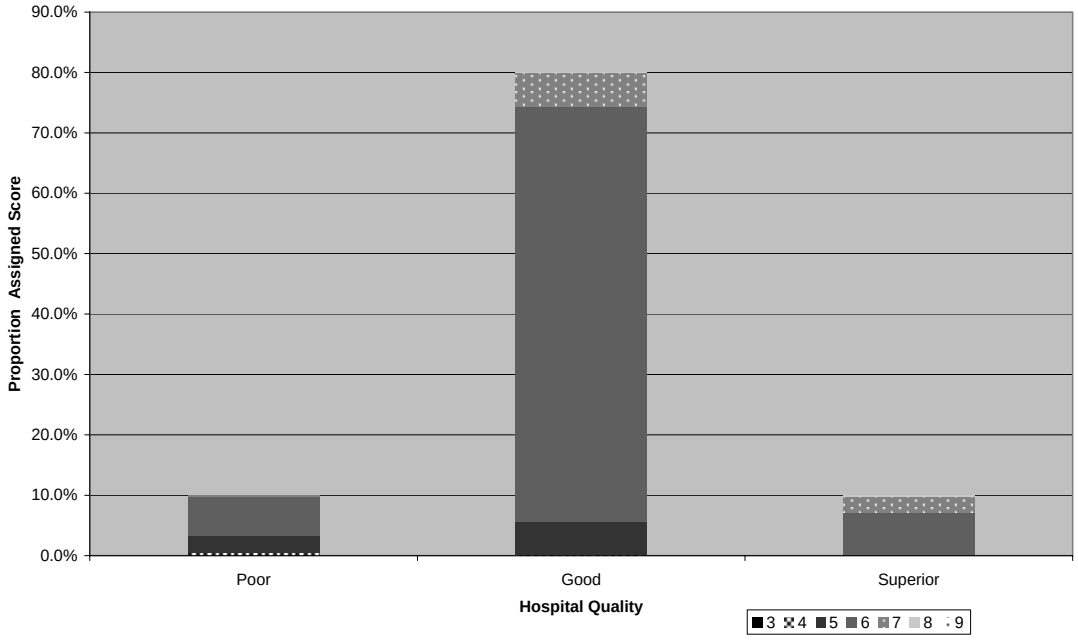
Figure C 9: Scenario 2: Proportion of Poor, Good, and Superior Hospitals with Each Type of Derivative Score



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

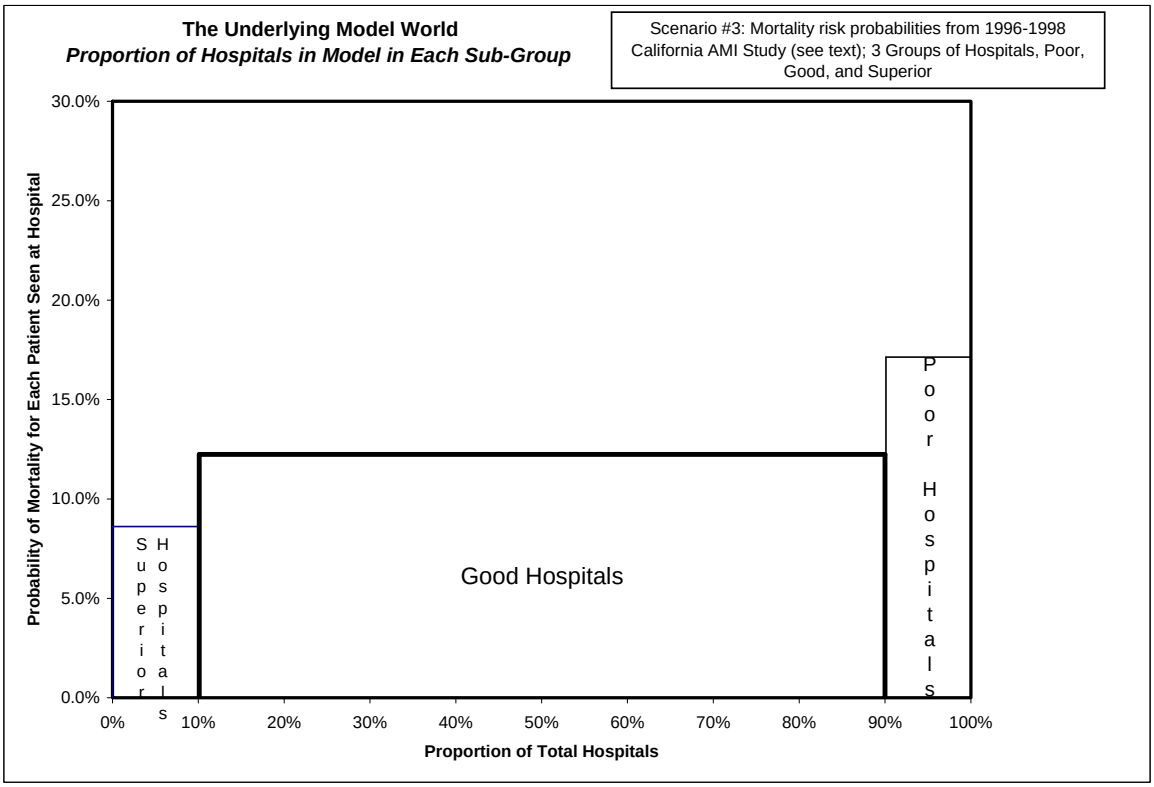
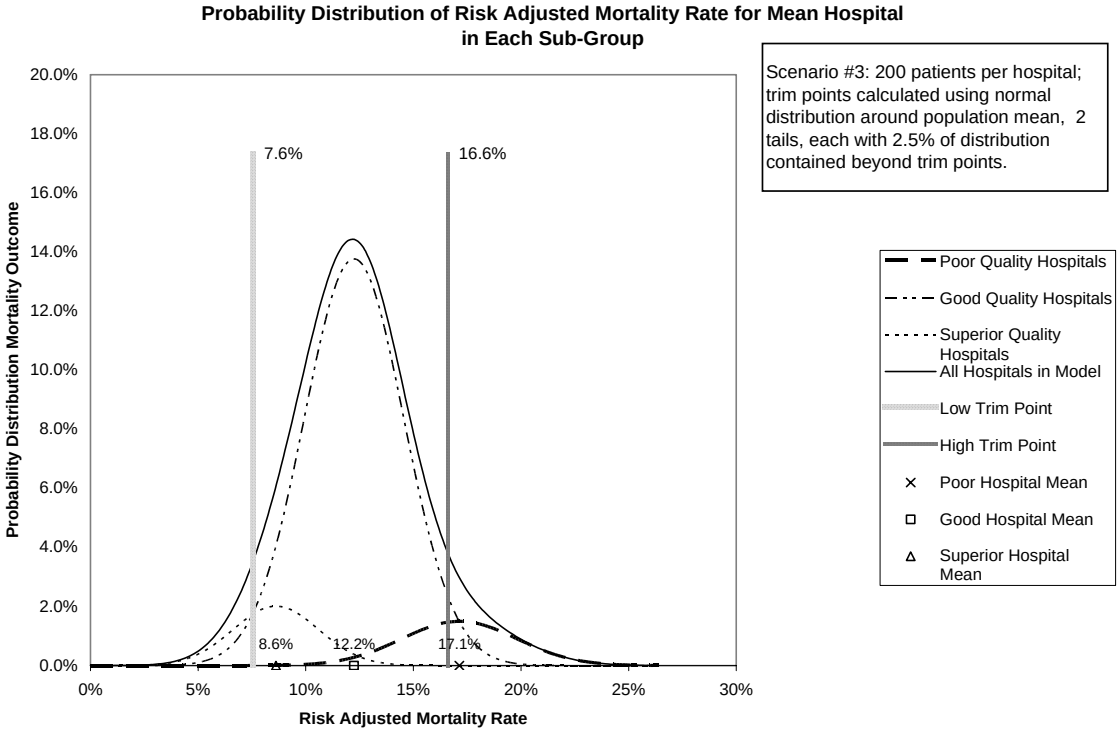


Figure C 12: Scenario 3: Hypothetical World and Evaluation Function



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

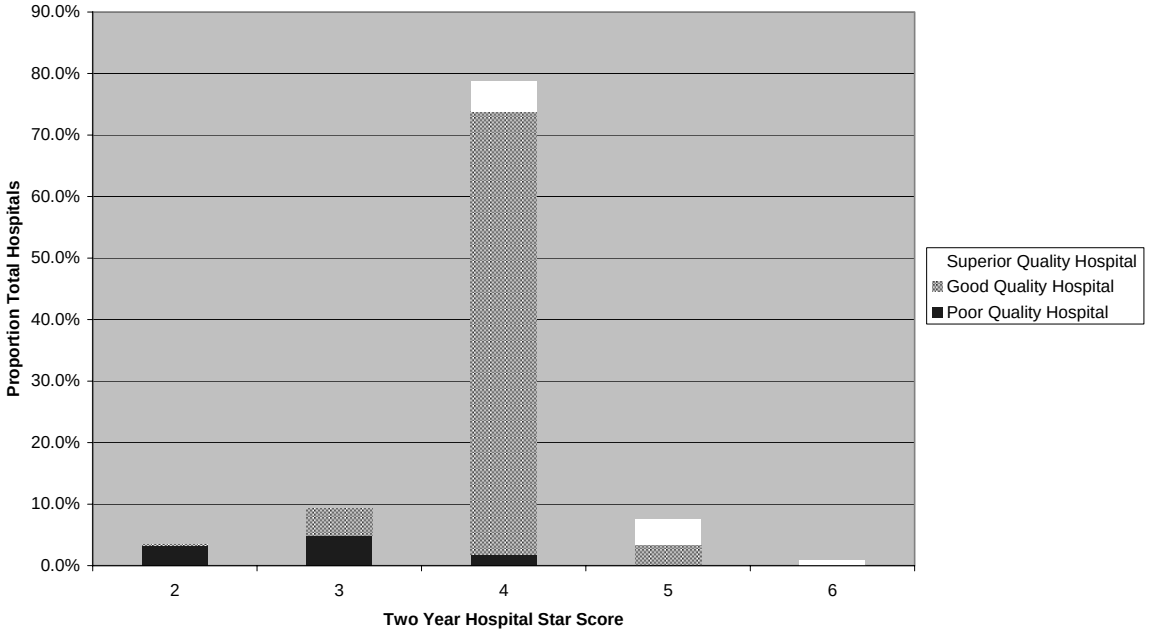
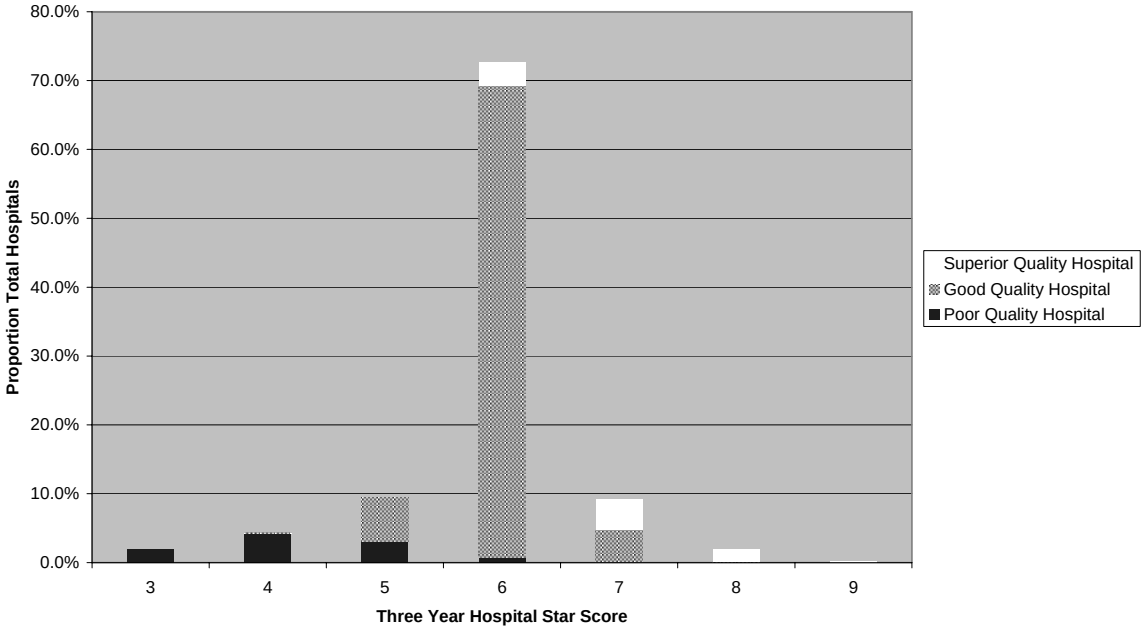
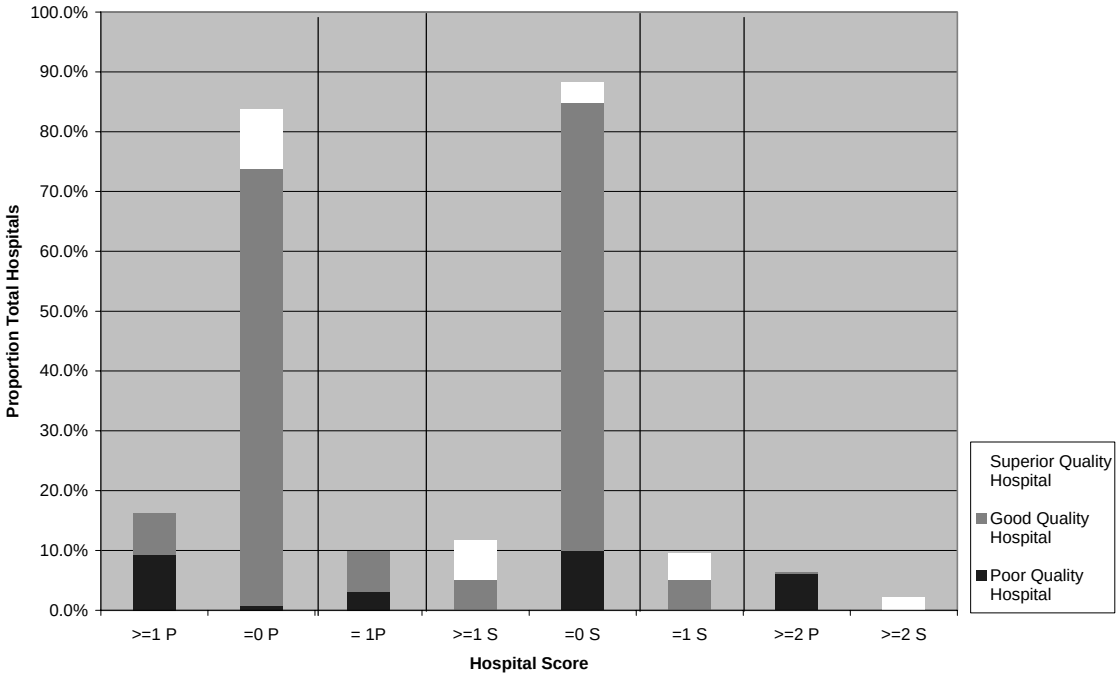


Figure C 14: Scenario 3: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score



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

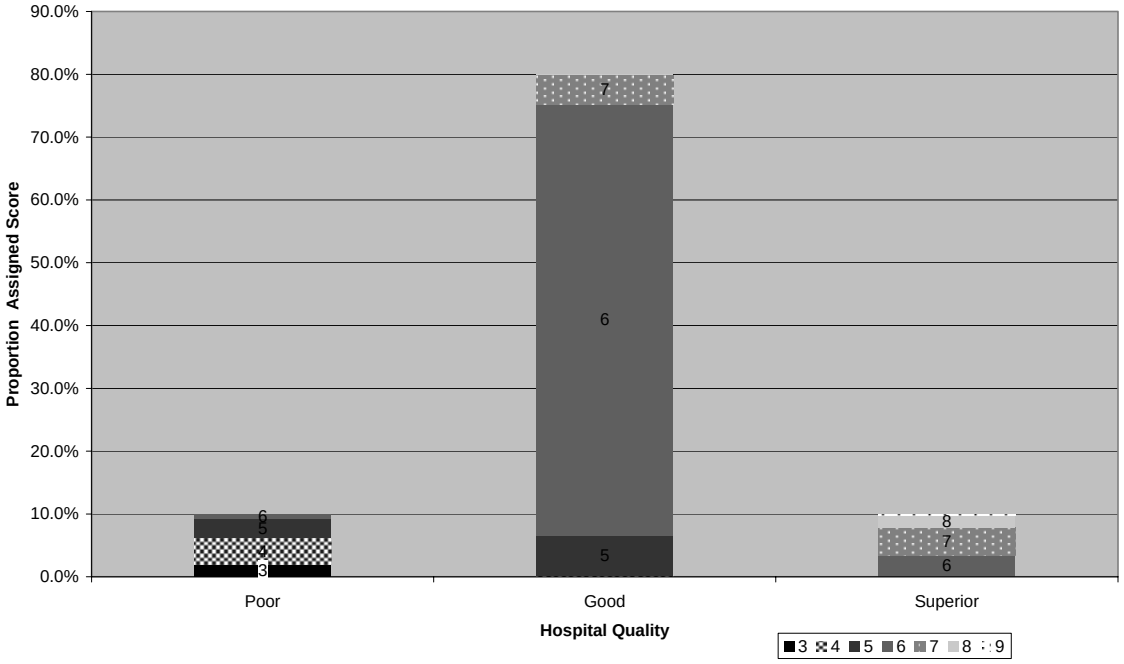


Figure C 17: Scenario 3: Expected Distribution of 3-Year Star Scores by Hospital Type

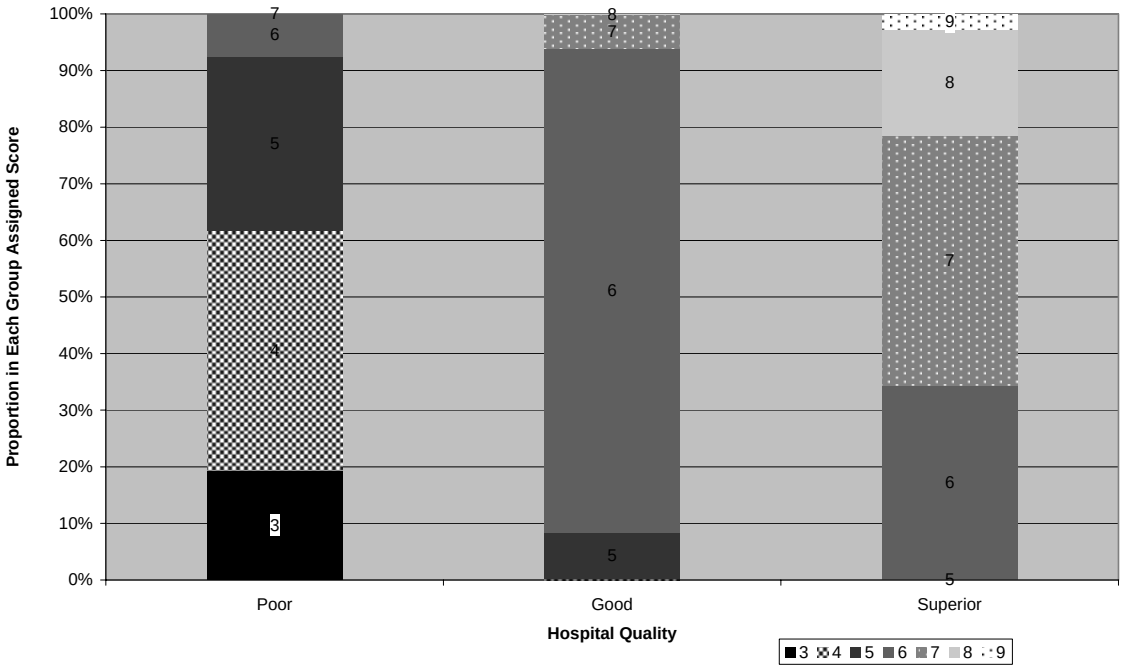
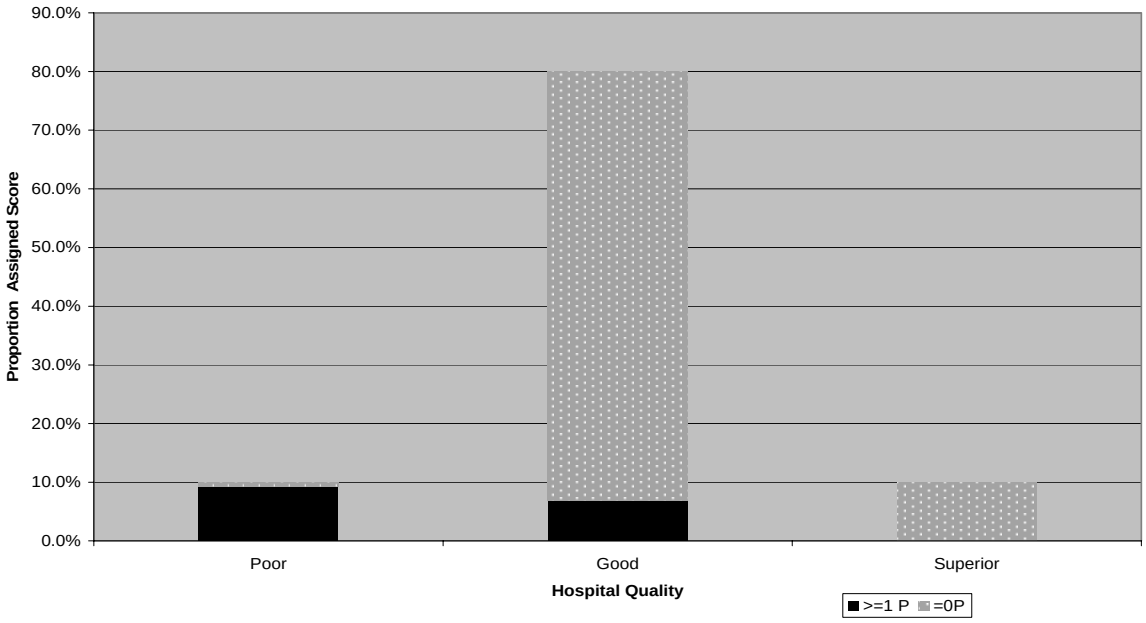


Figure C 18: Scenario 3: In 3 Years Ever Graded Poor vs. Never Graded Poor



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

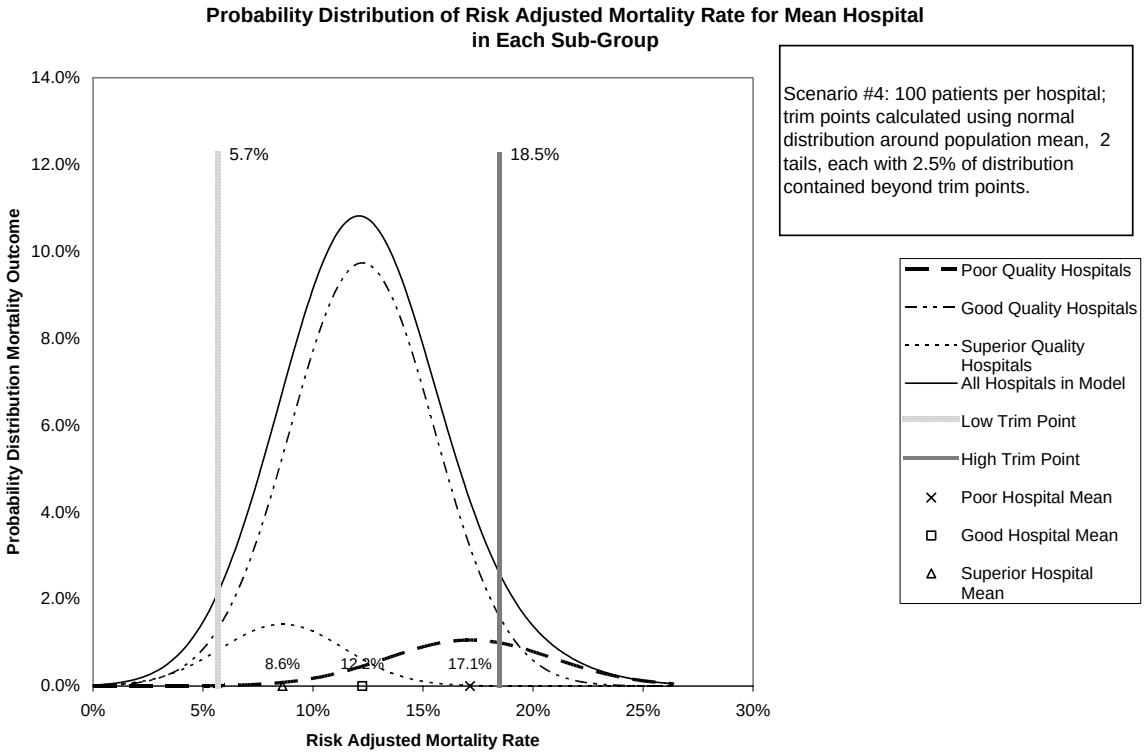
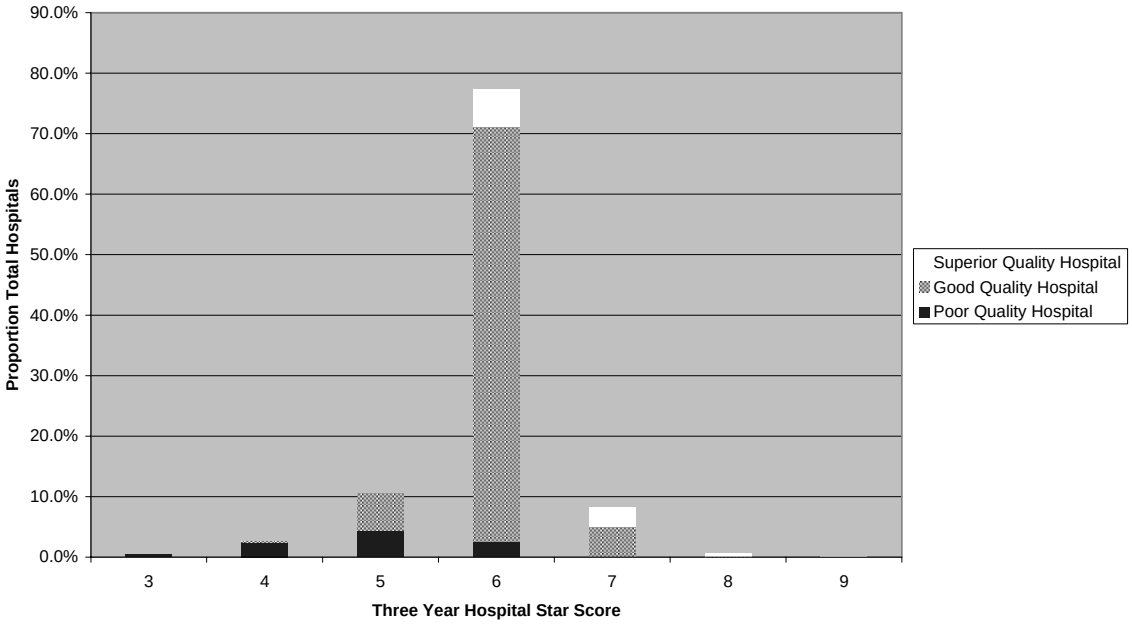
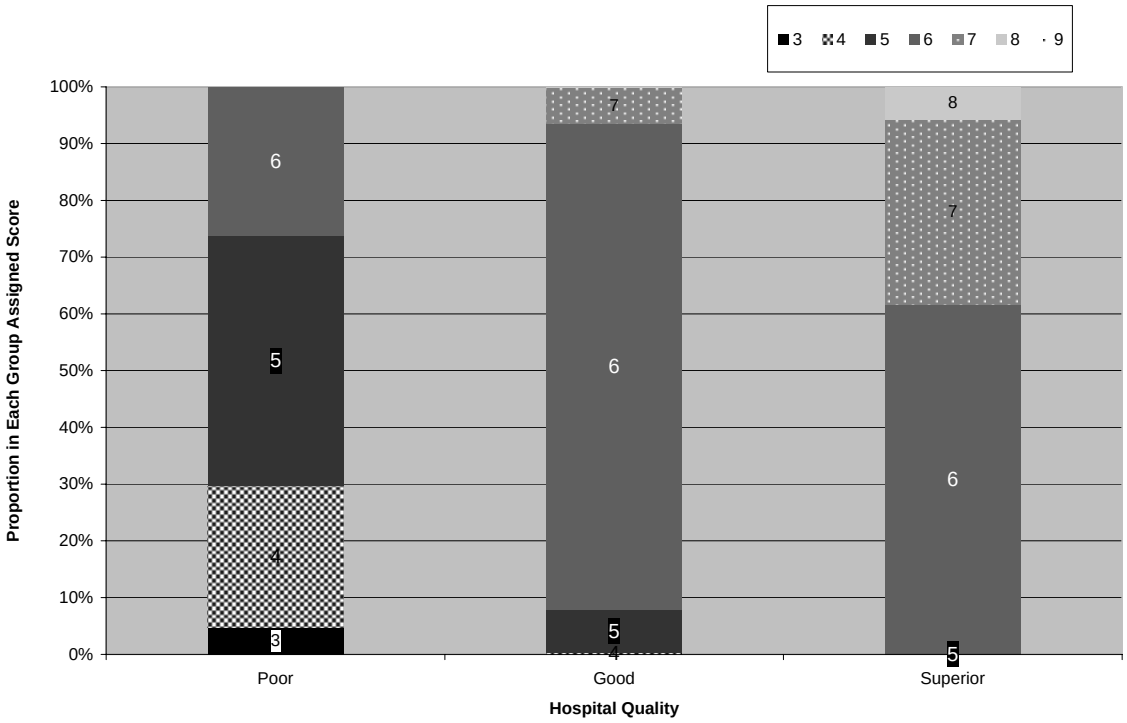


Figure C 20: Scenario 4: Proportion of Superior, Good, and Poor Hospitals by 3-Year Star Score



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

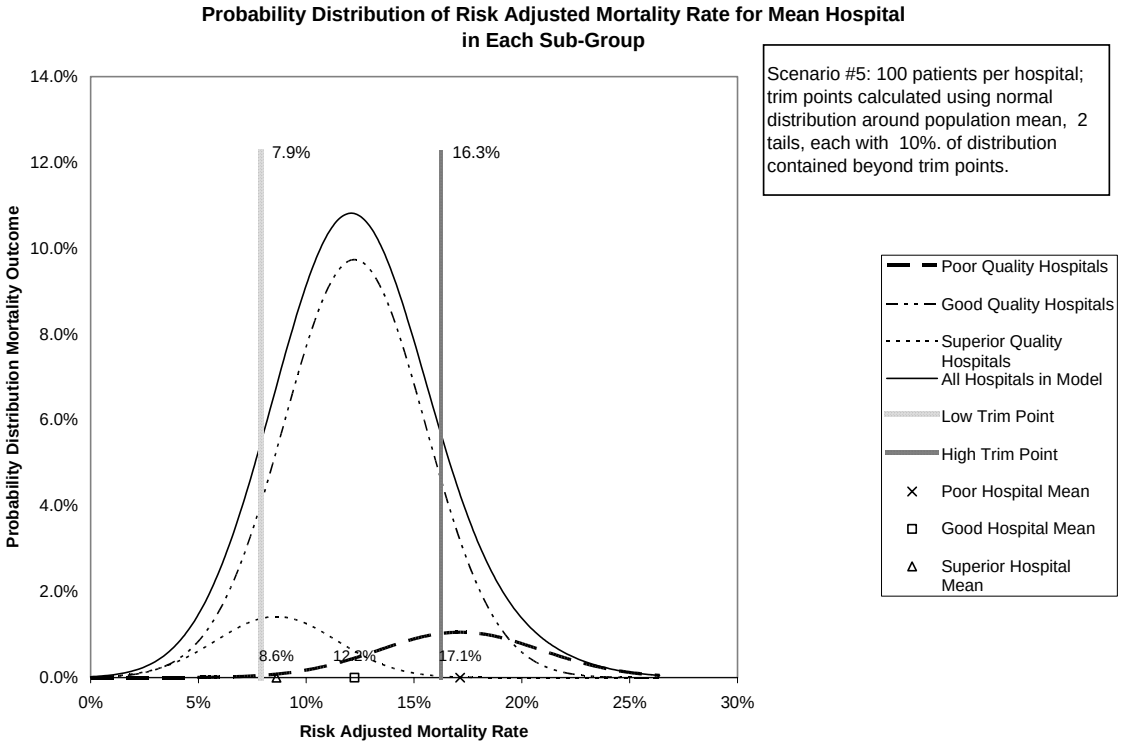
Figure C 21: Scenario 4: Expected Distribution of 3-Year Star Scores by Hospital Type



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

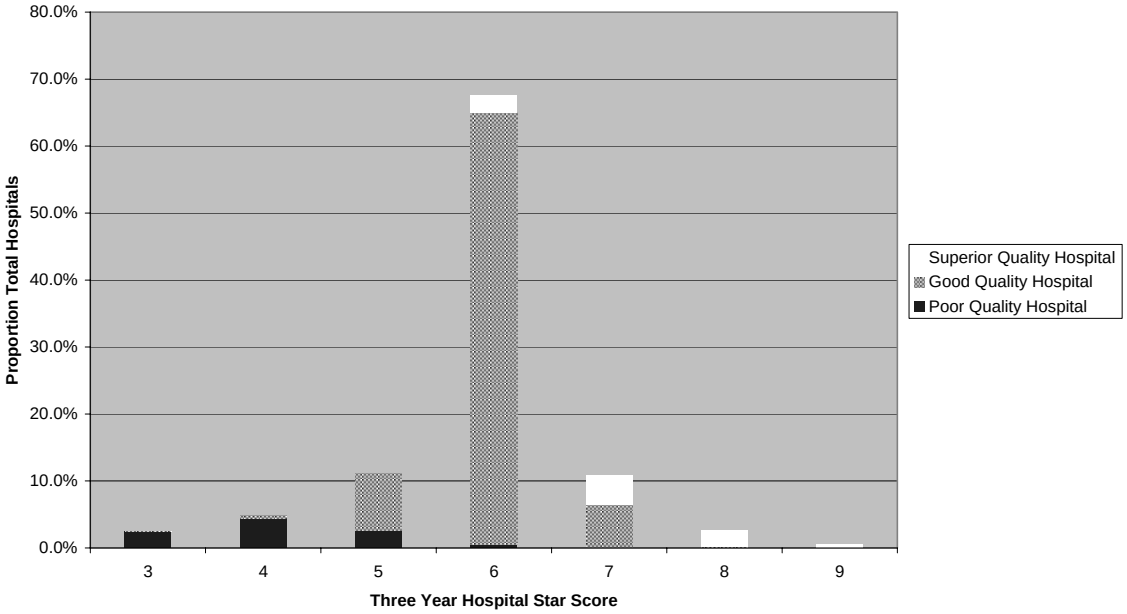
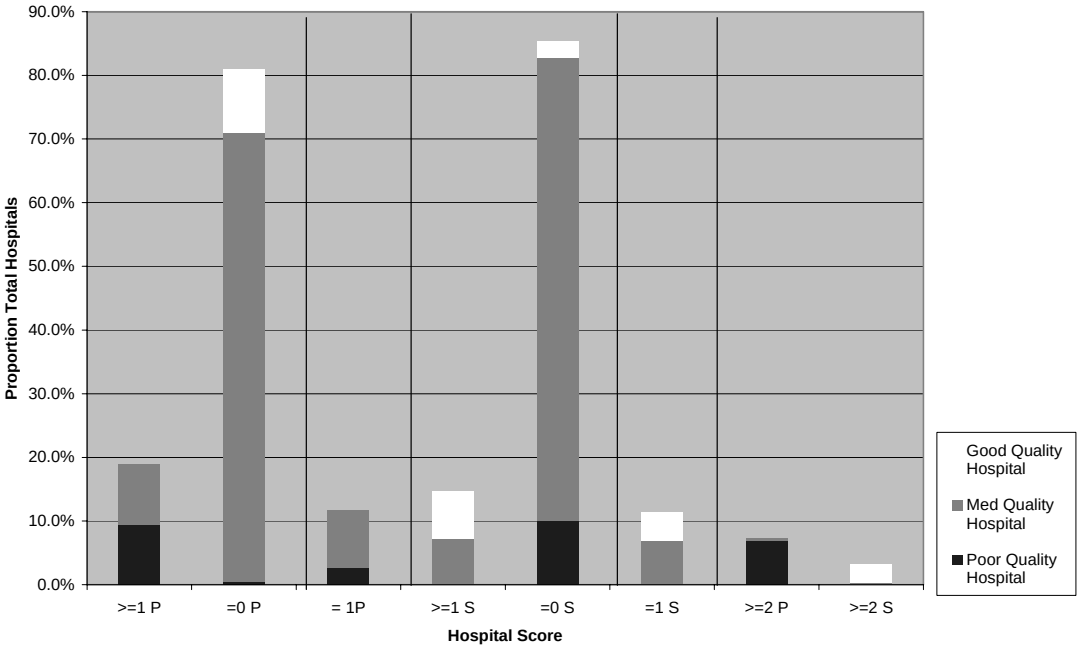
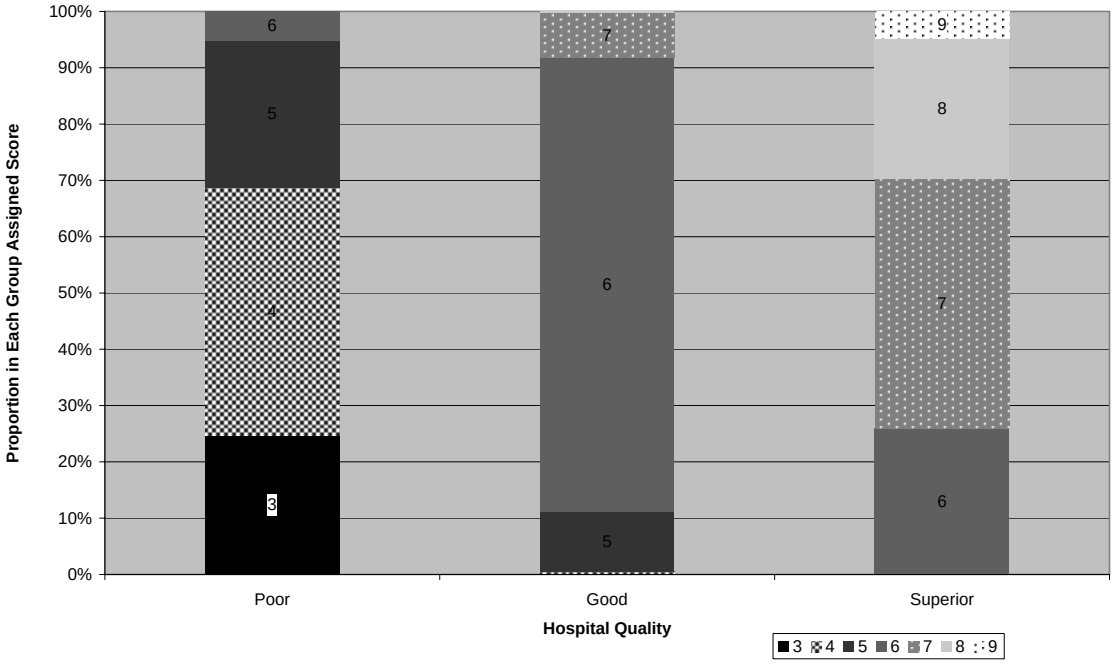


Figure C 24: Scenario 5: Three-Year Derivative Scores, Predictive Values



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

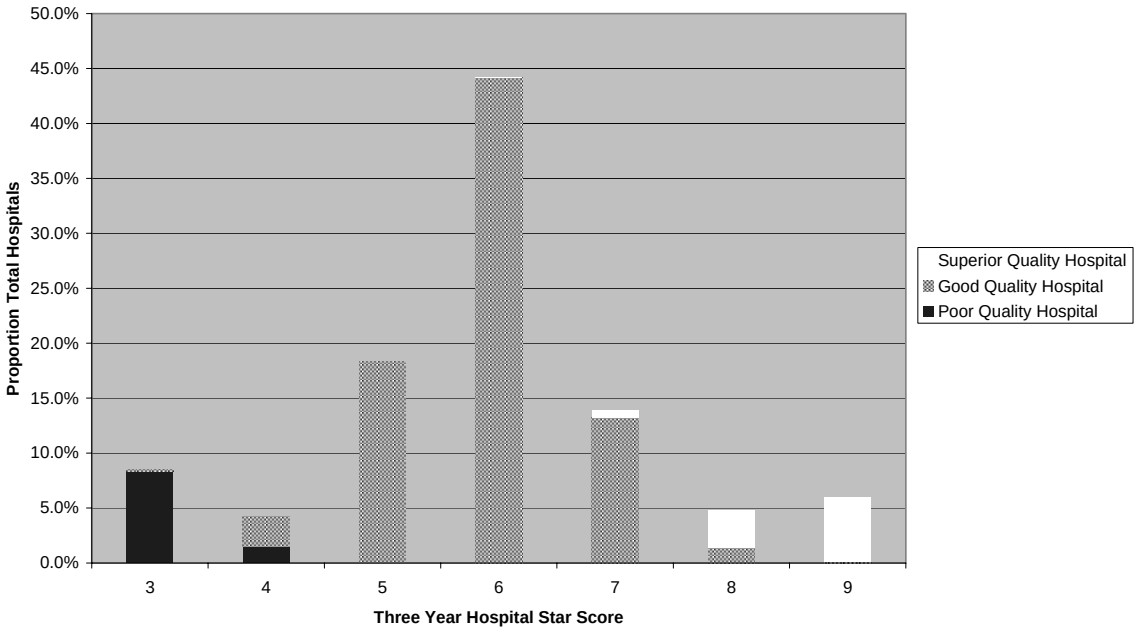


Figure C 27: Scenario 6: Three-Year Derivative Score Predictive Values

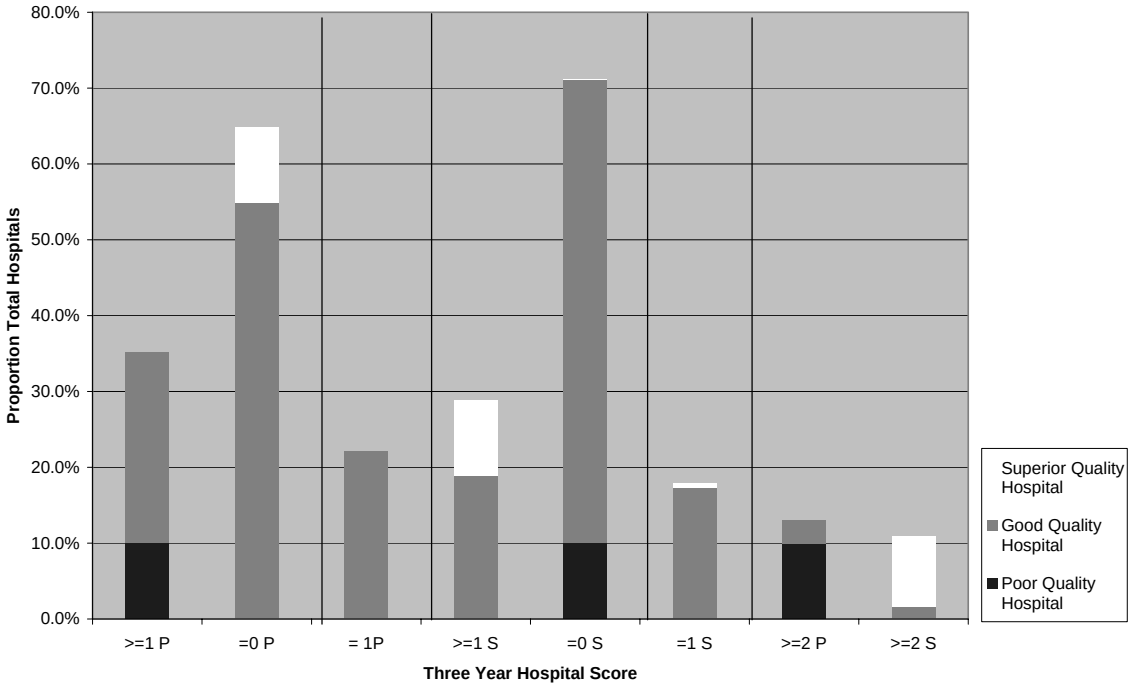


Figure C 28: Scenario 6: Expected Distribution of 3-Year Star Scores by Hospital Type

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.