

## **IMPROVING PROVIDER PROFILES FOR ASSESSING QUALITY OF CARE IN A HIGH RISK PATIENT POPULATION**

Douglas O. Staiger, Ph.D., Jeannette A. Rogowski, Ph.D., Jeffrey D. Horbar, M.D.,  
Michael Kenny, M.A., Jeffrey Geppert M.A., and Mark McClellan, M.D., Ph.D.

Corresponding Author  
Douglas Staiger, Ph.D.  
Department of Economics  
Dartmouth College  
Hanover, NH 03755  
Phone: (603) 646-2979  
Fax: (603) 646-2122  
Email: [doug.staiger@dartmouth.edu](mailto:doug.staiger@dartmouth.edu)

Word Count: 5752

Author Affiliations: Department of Economics, Dartmouth College, Hanover, New Hampshire and the National Bureau of Economic Research (Dr. Staiger), RAND, Arlington Virginia (Dr. Rogowski), Department of Pediatrics, University of Vermont College of Medicine and Vermont Oxford Network (Dr. Horbar ) Biometry Division University of Vermont (Mr. Kenny), Departments of Economics and Medicine, Stanford University, Stanford, California and the National Bureau of Economic Research (Dr. McClellan), National Bureau of Economic Research (Mr. Geppert).

This research was funded by grant number R01 HS10328 from the Agency for Healthcare Research and Quality.

**Objective** To evaluate a new type of performance measure that combines a broad range of information on providers, and compare the reliability, stability, and bias of alternative performance measures, with an application to the care of infants with very low birth weights.

**Data Sources/Setting** A total of 94,222 very low birth weight infants (under 1500 g) cared for by 317 NICUS that belonged to the Vermont Oxford Network between 1994 and 1999.

**Design** Performance measures were constructed for mortality (3-day, 28-day, in-hospital) and major comorbidities (nosocomial infection, severe intra-ventricular hemorrhage). Detailed case-mix adjustment was performed using patient characteristics available immediately after birth.

**Principal Findings** Compared to conventional risk-adjusted outcome measures, the reliability of provider profiles increased substantially when performance measures were constructed using methods that integrate information across measures and over time. The reliability of mortality estimates more than doubled for NICUs treating less than 40 very low birth weight infants each year, while standard errors on these estimates fell more than in half for most providers. The resulting performance measures were highly correlated from year to year (correlation=0.89 for 28-day risk adjusted mortality,  $p<.001$ ) and were able to prospectively identify NICUs with a 69% 28-day mortality difference in 1999 (13.5% versus 8.0%,  $p<.001$ ). Inadequate case mix adjustment resulted in the performance measures of the best units being biased towards poorer performance, while in-hospital mortality measures biased the estimates for NICUs with high transfer rates toward understating their mortality.

**Conclusions** Performance measures for very low birth weight infants reliably estimated large and stable differences in patient outcomes across a national sample of NICUs, when developed in conjunction with statistical methods that integrated performance measures and that used detailed data for measuring patient outcomes and risk factors.

**Keywords** Provider profiling, hospital performance, quality of health care, low birth weight infants, newborn intensive care units.

## INTRODUCTION

In recent years there has been an increased emphasis on improving the quality of medical care and on holding health care providers and health plans accountable for the quality and cost of the services they provide. There has been a wide array of public and private efforts at creating performance measures based on adverse events, such as mortality or complication rates, or based on inappropriate or avoidable use of medical care, such as rates of cesarean section or low birth weight (Davies and Marshall 1999; Marshall, Shekelle et al. 2000). At the federal level, the Health Care Financing Administration and the Agency for Healthcare Research and Quality have both developed hospital performance measures (AHCPR 1994; Mennemeyer, Morrisey et al. 1997; AHRQ 2001), as has the National Health System in England (Harris 2001). States such as New York and Pennsylvania publish report cards for both physicians and hospitals based on mortality rates for CABG surgery (Hannan, Kilburn et al. 1994; Bentley and Nash 1998). Consortiums such as the Vermont Oxford Network and the Northern New England Cardiovascular Study Group use provider profiles as part of their quality improvement efforts (O'Connor, Plume et al. 1996; Horbar 1999). Measurement systems for profiling health plans have also evolved, such as the Health Plan Employer Data and Information Set (HEDIS) and Consumer Assessment of Health Plans (CAHPS) (Epstein 1998).

This growth in the uses of provider profiling has occurred despite many negative and cautionary evaluations of typical performance measures. A number of studies have documented the instability of performance measures from one year to the next, with top ranked providers often appearing below average the next year (Parry, Gould et al. 1998; McClellan and Staiger 2000). There is a large related literature documenting the imprecision in many performance measures, particularly when there are relatively few observations per provider, when adverse

outcome rates are relatively low, and when providers have little control over patient outcomes or there is little variation in important processes of care.<sup>1</sup> Finally, in addition to these problems with the reliability of the measures themselves, many studies have documented that case-mix differences across providers lead to important biases in the absence of extensive risk-adjustment, although this is not always the case.<sup>2</sup> Perhaps not surprisingly, most providers view performance measures with considerable skepticism (Berwick and Wald 1990; Davies and Lampel 1998) and reports of provider mortality rates appear to have little effect on where patients seek their care (Schneider and Epstein 1996; Hibbard and Jewett 1997; Hibbard, Jewett et al. 1997; Menemeyer, Morrissey et al. 1997; Schneider and Epstein 1998).

These perceived problems with currently available provider performance measures have resulted in considerable debate over the appropriate uses of such information. But given the growing pressures placed on health care providers to provide efficient high quality care, the trend toward greater accountability is likely to continue – with increasing pressure to develop and apply measures to finer levels of analysis (e.g. individual physicians, narrow groups of patients). The fundamental question, therefore, is whether we can develop better methods for comparing performance across providers in such situations.

We examine these issues in the context of a high-risk patient population, infants with very low birth weights (under 1500 grams). We evaluate a new method of estimating provider

---

<sup>1</sup> See for example: (Park, Brook et al. 1990; Thomas, Holloway et al. 1993; Hofer and Hayward 1995; Hofer and Hayward 1996; Thomas 1996; Localio, Hamory et al. 1997; Normand, Glickman et al. 1997; Parry, Gould et al. 1998; Thomas and Hofer 1998; Hofer, Hayward et al. 1999; Thomas and Hofer 1999).

<sup>2</sup> Examples of studies finding bias include: (Iezzoni, Ash et al. 1995; Iezzoni, Shwartz et al. 1995; Iezzoni, Ash et al. 1996; Iezzoni, Shwartz et al. 1996; Iezzoni, Shwartz et al. 1996; Hannan, Racz et al. 1997). Examples of studies finding no bias include: (Luft, Romano et al. 1996; Iezzoni, Ash et al. 1997; Pine, Norusis et al. 1997; Krumholz, Chen et al. 1999).

performance (McClellan and Staiger 1999; McClellan and Staiger 2000), which combines information from many sources to address the problems just described. Infants with very low birth weights are an ideal population to study for a number of reasons. Very low birth weight infants suffer from high levels of adverse events thought to be potentially avoidable, with mortality in the first year of fifteen percent accounting for 51 percent of infant deaths (Guyer, MacDorman et al. 1998). Yet there is no consensus on how to measure quality of medical care for very low birth weight infants, and measures based on patient outcomes have been found to be both unstable and unreliable because of the small numbers of such infants treated in a typical hospital (Parry, Gould et al. 1998).

This study was performed in collaboration with the Vermont Oxford Network as part of their ongoing quality improvement activities. Thus, we build on an existing base of practical experience, with data and performance measures that have been developed and refined over the last decade. The Vermont Oxford Network was established in 1989 with the goal of improving the quality and safety of medical care for newborn infants and their families through a coordinated program of research, education and quality improvement projects. The Network is a private, non-profit corporation supported by membership fees, grants, and contracts. In support of its mission the Network maintains a Database for infants 401 to 1500 grams cared for at participating centers. Based on these data, the Network routinely provides its members with individual performance reports and also supports several collaborative quality improvement efforts among its members (Horbar and al. 2001). One goal of this study is to provide better methods for the reporting of provider performance to NICUs that are members of the Network, to assist them in understanding and identifying opportunities for improvement.

## **METHODS**

### *Sites and Patient Sample*

Data for this study come from the Vermont Oxford Network for the years 1994-1999. The Vermont Oxford Network is a voluntary, collaborative network of neonatal intensive care units with over 380 institutions now contributing to the Network Database. The Network Database contains detailed, uniform clinical and treatment information on half of all very low birth weight infants born in the United States each year.

The study included 317 NICUs from the Vermont Oxford Network, of which 90 had data for all years and the remainder had data for 1-5 years. Hospitals with less than a full panel of data were mainly comprised of those that joined the Network after 1994. For the purpose of creating performance measures, each very low birth weight infant was assigned to the VON NICU to which it was first admitted. We created performance measures for any year in which a NICU admitted at least 5 infants with very low birth weights. Overall, over the six-year period of the study, these units cared for 94,222 infants with very low birth weights. The average number of very low birth weight infants treated in any given year in a unit was 79. This ranged from 5 to 288 across units.

### *Variables*

All variables for the analyses were obtained from the Vermont Oxford Network Database. The database contains information on all infants with birth weights of 401 to 1500 grams born at member institutions or admitted to them within 28 days of birth. This includes infants who die in the delivery room or other locations in the hospital even if they are not admitted to the NICU. The Network tracks all transfer hospitalizations until the child is either discharged home or dies.

Key outcome measures include mortality at 3 and 28-days after birth, and overall in-hospital mortality, including all transfer hospitalizations. For some analyses, we restrict to mortality at the admitting hospital (not including transfers) within 28 days. In addition, two major comorbidities were considered: nosocomial infection and severe intra-ventricular hemorrhage. Nosocomial infection is defined as the presence of either bacterial sepsis or meningitis or a coagulase negative staphylococcal infection more than 3 days after birth. Intra-ventricular hemorrhages are defined to be severe if they are of grade 3 or higher (Papile, Burstein et al. 1978).

The Vermont Oxford Network Database contains detailed information on each infant at the time of birth that we use for the purposes of risk adjustment. The variables that we use to adjust for severity include: birth weight, gestational age, major birth defect, small size for gestational age, multiple birth, 1-minute Apgar score, sex, race (Hispanic, white, black or other), delivery by cesarean section, location of birth (inborn/outborn), prenatal care, and early bacterial sepsis (within 3 days of birth and, therefore, likely to be pre-existing). With the exception of birth weight and sex (which were never missing), missing values for all variables were imputed and a corresponding variable was created indicating that the data were imputed. Less than two percent of infants had missing data for any given variable.

### *Constructing Conventional Risk-Adjusted Outcomes*

Conventional risk-adjusted outcome rates for each of the mortality and complication outcomes discussed above are constructed for each NICU based on the Network risk-adjustment model (Horbar 1999). Separate models are estimated for each year of data, 1994 through 1999. We have augmented the list of covariates used in the usual Network model to include measures

of birth weight, prenatal care, early bacterial sepsis, and indicators for when each variable has been imputed. This risk-adjustment model has been used successfully by the Network to adjust for differences in case mix among hospitals (Horbar 1999) with a Hosmer-Lemeshow Goodness-of-Fit  $p=0.79$  and area under the Receiver Operating Curve of 0.88.

Risk-adjusted outcome rates are estimated using a linear regression model. In particular, for each year we construct a risk-adjusted outcome rate for each NICU from a patient level regression of the outcome variable (e.g. a dummy variable indicating if the infant was dead at 28 days) on patient covariates and NICU-specific intercepts. The resulting NICU-specific intercepts are the estimates of risk-adjusted outcome rates for each NICU, and the standard errors on these estimates can be used to form confidence intervals. These risk-adjusted outcome rates are estimates of the expected outcome rate at each NICU for an average patient. In the technical appendix we discuss the reasons for estimating a linear regression specification rather than a logistic regression (which is more commonly used for risk-adjustment). However, NICU-specific estimates of risk-adjusted outcome rates are highly correlated from the two methods.

### *Estimating Variance Across Units, Correlation over time, and Reliability of Performance Measures*

The usual approach to assessing the magnitude of differences in performance across units and the stability of these differences over time is to calculate statistics such as the standard deviation across units and the correlation between years based on conventional estimates of risk-adjusted outcome rates for each unit. A large standard deviation indicates large differences in performance between units, while a high degree of correlation from one year to the next indicates stability of performance over time. However, conventional risk-adjusted outcome rates are based

on small numbers of patients and therefore are noisy measures of true performance. Because they have substantial random variation, they tend to overstate the amount of true performance variation across units in any particular measure in any given year. This random variation or measurement error in conventional outcome measures also tends to obscure the true correlation in performance between years: even when true performance does not change, the noise in measured performance generates random variations from one year to the next. Conventional measures are therefore typically unstable.

Our method corrects for the random noise (due to estimation error) in conventional outcome rates in a straightforward manner (McClellan and Staiger 1999; McClellan and Staiger 2000). The observed variance in conventional outcome measures across units is composed of signal variance (real performance differences across units) plus noise variance (estimation error in conventional outcome rates). Analyses should be based on signal (true) variance only. To estimate the signal variance for each performance measure, we subtract an estimate of the noise variance from the total variance observed in each measure (which reflects both signal and noise variance). The noise variance is estimated based on the number of patients seen at each hospital. In other words, we estimate the amount of true variation in performance based on how much the observed variation exceeded what would have been expected due to sampling error alone. We then use these estimates of the signal variance to calculate corrected statistics such as the standard deviation across units and the correlation in performance between years.

Rather than calculate separate estimates of the variance and correlations in performance for each year (which would be a total of 21 estimates in our 6 years of data), we summarize these relationships with a simple autoregressive structure (Greene 1997). This structure assumes that the variance in true performance across units is the same in each year, and that the correlation in

true performance between any two years is stable over time and declines geometrically in proportion to the length of time between years. These assumptions cannot be rejected ( $p > .10$ ) for any of the outcome measures used in our analysis. One advantage of this model is that it allows one to summarize the performance data with two simple parameters: the variation across units in true performance, and the correlation between one year and the next. A second advantage of this model is that it allows one to estimate the expected correlation between current and future levels of performance, which is necessary if one wishes to forecast future performance levels. A useful test of the adequacy of this model is its ability to forecast future performance. For details on estimation, see the technical appendix.

Finally, a common metric used to evaluate and compare performance measures is reliability. Reliability is the proportion of the observed variation in a performance measure that is due to true differences in performance across providers, as opposed to chance variation. Equivalently, a measure's reliability is an estimate of what proportion of the real variation in performance could be predicted by the observed measure (similar to an R-squared value in a regression). We calculate the reliability of conventional performance measures using the ratio of signal variance to signal plus noise variance, where the signal and noise variance are estimated as previously discussed. Performance measures are less reliable for providers with fewer patients because these measures have more noise variance, e.g. are estimated with more error.

### *Filtering the Risk-Adjusted Outcomes*

Conventional estimates of risk-adjusted outcome rates at any given hospital in a given year can be both unstable and unreliable, because of the typically small samples of patients treated at a given hospital. We construct alternative outcome measures using a multivariate

empirical Bayes approach that has recently been used in a variety of applications to improve the reliability and stability of performance measures (McClellan and Staiger 1999; McClellan and Staiger 2000; McClellan and Staiger 2000; AHRQ 2001; Kane and Staiger 2001; Kane and Staiger 2002 (forthcoming)). Our methodology uses an estimation technique that is closely related to hierarchical Bayes methods (Normand, Glickman et al. 1997; Burgess, Christiansen et al. 2000), but permits rapid calculation of results for large datasets.

The methods we use are a generalization of the idea of applying a “shrinkage factor” to each hospital’s estimate, so that less reliable estimates (those with a lower signal to noise ratio) are shrunk toward the overall average (O’Hagan 1994; Hofer, Hayward et al. 1999). In addition to shrinking any single year’s estimate toward an overall average, multivariate methods also account for correlations in performance over time by averaging estimates over years in which true performance was likely to be similar. This method smoothes performance estimates over time by creating a sophisticated moving average of each provider’s conventional risk-adjusted outcome rates, thus eliminating much of the excess volatility arising from noise in the conventional measures. We refer to the resulting estimates as “filtered” estimates, because these estimates eliminate much of the noise in the original data.

More specifically, the filtered performance estimate for a given hospital in a given year is a weighted average of average performance among all hospitals in that year, and conventional performance estimates for that hospital from every available year. The weights depend on only four factors. The weight placed on a hospital’s own performance (as opposed to average performance at all hospitals) increases with the number of patients seen at the hospital and with the amount of true variation in performance across all the hospitals. The weight placed on performance estimates from other years increase with the nearness in time to the year of interest,

and with the correlation in performance from one year to the next. In constructing the weights, we use estimates of the true variation and correlation in performance, derived as previously discussed. Standard errors for the filtered measures will be smaller, and reliability higher, when the correlation between years is larger or more years of data are available. For details on how the weights and other statistics are constructed, see the technical appendix.

## RESULTS

### *Stability of Performance Measures Over Time*

Table 1 reports estimates of the mean and standard deviation of various performance measures across NICUs in 1999, along with the correlation in these performance measures between 1998 and 1999. The performance measures include risk-adjusted mortality rates at 3 days, 28 days, and overall, and risk-adjusted rates of nosocomial infection and severe intraventricular hemorrhage. Estimates of the standard deviation in performance across NICUs and of the correlation in performance between 1998 and 1999 are reported in the second and third column of Table 1. Because of the estimation error in each year's performance measures, conventional estimates will tend to overstate the amount of true performance variation in any one year and understate the true correlation in performance between years. The final two columns report estimates that correct for the presence of estimation error as discussed in the methods section.

Table 1 illustrates the importance of accounting for estimation error in any comparison of performance measures across providers. The performance estimates in Table 1 that do not account for random variation overstate the standard deviation (variation) in performance across NICUs by as much as a factor of two, and understate the year-to-year correlation by a similar

proportion. For example, the conventionally estimated standard deviation across NICUs in 28-day mortality is over 40% of the average mortality rate (5 percentage points relative to an average rate of 11.8 percent), and the estimated correlation from 1998 to 1999 is only 0.27. Thus, the conventional estimates suggest that there are substantial differences across NICUs in mortality rates, and that these differences are not very stable over time. In contrast, estimates that correct for estimation error suggest that the standard deviation in 28-day mortality across NICUs is less (2.5 percentage points), but these differences are very persistent between 1998 and 1999 with a correlation of 0.89.

After correcting for estimation error there remain large and stable differences in patient outcomes across NICUs. For example, 3-day, 28-day and in-hospital mortality measures have standard deviations across units of 1.9 (with an average mortality rate of 6.1 percent), 2.5 (average mortality 11.8 percent) and 3.0 percentage points (average mortality 14.2 percent) respectively ( $p < 0.001$  in all cases). Variation in outcomes across hospitals not due to chance is similarly large for major morbidities. The standard deviation across units is one third of the average rate for nosocomial infection and one fifth of the average rate for intraventricular hemorrhage. These differences across units are also quite stable over time, with the estimated correlation in these measures from year to year ranging from .85 to 1.0.

The consistency of the performance of providers from year to year, after accounting for random chance, implies that the reliability and stability of performance measures can be increased substantially by pooling information on performance across years. Figure 1 demonstrates the effects of using the filtering methods on case-mix adjusted 28-day mortality rates for 9 NICUs. These NICUs have data from all years of the study (1994-1999), with mortality rates based on between 45 and 50 admissions per year. Each NICU is labeled with its

overall rank in the Network from 1999, based on filtered 28-day mortality, with 1 being the lowest mortality in the Network and 297 being the highest mortality.

The conventional case-mix adjusted estimates (plotted in the left-hand panel) are very unstable, with large differences across providers at any point in time and large jumps both up and down over time. This instability is what makes conventional estimates difficult to use for the purposes of measuring and monitoring quality. Much of the lack of stability comes from the imprecision in conventional estimates, that is, large standard errors. For conventional estimates standard errors averaged 4 percentage points. In contrast, the filtered estimates are more precisely estimated, with standard errors of less than 2 percentage points, and are considerably more stable because of the smoothing that is done across years. The right-hand panel of Figure 1 shows the case-mix adjusted mortality rates after filtering. The filtered estimates are both able to more clearly differentiate between good and bad performance, and also make it easier to observe when NICU performance is improving. For example, in the filtered estimates it is apparent that hospital 287 has relatively high mortality throughout this period, while this fact was not obvious in the conventional estimates.

The stability of the filtered estimates from year to year is a direct result of the estimates from Table 1, which indicates that true NICU performance is stable over time once random estimation error is filtered out. Thus, while conventional measures are only weakly correlated over time, the filtered measures are strongly correlated over time. This difference between conventional and filtered estimates is further illustrated in Figure 2. This Figure plots risk-adjusted 28-day mortality measures from 1998 against these same measures in 1999, with conventional measures displayed in the left hand panel and filtered measures in the right hand panel. The conventional measures show little correlation, due to the large amount of noise in the

data. The filtered measures more clearly represent the persistent, underlying performance differences across units.

More reliable performance measures enable more accurate forecasting of future performance. Table 2 demonstrates that the filtered performance measures can accurately forecast large differences in patient outcomes two years into the future. Such an ability to identify best and worst performers based on historical data is an important component of benchmarking efforts. Filtered estimates of 28-day mortality rates were constructed using four years of data from 1994 to 1997, and then used to identify the top and bottom 10 percent of providers (along with the remaining 80 percent). In Table 2, we evaluate the actual mortality performance of these providers two years later (in 1999). As can be seen in the first row of the table, actual adjusted 28-day mortality rates in 1999 were 69% higher among NICUs identified as high mortality in 1997 as compared to NICUs identified as low mortality in 1997 (13.5% versus 8.0%,  $p < 0.001$ ). Moreover, the filtered estimates from 1997 (14.5% versus 9.2%) were accurate predictions of the actual mortality rates observed in each group in 1999. In each case the forecasts are within one standard error of the actual mortality rates observed in 1999. In contrast, conventional estimates of mortality rates from 1997 (which would typically be used to identify top and bottom performers in the absence of filtering) significantly overstate the difference in mortality rates between the top and bottom performers ( $p < 0.001$ ). This is because they rely too much on actual (noisy) results in 1997.

### *Reliability of Performance Measures*

A fundamental problem with conventional performance measures is that they are not reliable; that is, most of the apparent variation in the measures is due to chance rather than true

differences across providers. Figure 3 shows the estimated reliability – the share of the variation that can be attributed to true performance differences rather than random noise – of conventional and filtered estimates of 28-day mortality as a function of the number of patients admitted to a provider. The filtered estimates are always much more reliable, particularly for providers that treat small numbers of infants with very low birth weights each year. For example, for a median size NICU (with 66 VLBW infants per year) the reliability of the filtered estimates is around 60%, while the reliability of the conventional estimates is just below 40%. For a smaller provider, treating only 40 patients per year, filtering doubles the reliability of 28-day mortality estimates, from 25% for conventional estimates to 50% for filtered estimates. For these NICUs, small sample sizes generate high levels of random noise in performance measures, and thus it is difficult to discern the true, underlying quality of care from a single year of data.

Higher reliability of the filtered estimates translates into a better ability to discriminate performance between providers. This point is illustrated in Figure 4, which ranks the 90 NICUs that were present in the study for all 6 years based on risk-adjusted 28-day mortality rates in 1999. It is evident that the filtered estimates (right panel) are much more precisely estimated, as demonstrated by confidence intervals that are on average half as large as the confidence intervals for conventional measures (left panel). The confidence intervals in the figures are constructed so that any two providers whose confidence intervals (as measured by 1.4 standard errors) do not overlap are significantly different ( $p < 0.05$ ). Thus, based on the filtered estimates, we can identify 15 providers with intervals above the national average (marked H) that are significantly different from the 13 providers with confidence intervals below the national average (marked by L). Thus, these estimates are able to identify significant performance differences between the highest and lowest performing units. Further, the conventional estimates rank NICUs quite

differently, with many of the NICUs identified as the highest and lowest performing units based on filtered estimates appearing not to differ significantly from ordinary performance based on conventional estimates (and vice versa).

### *Bias in Performance Measures*

A second fundamental problem with provider profiles is that they can be biased if the underlying data systems used to create estimates are inadequate. In this section we explore the bias inherent in a limited ability to adjust for case mix adjustment and failure to account for patient transfers among infants with very low birth weights.

Inadequate risk adjustment will lead to bias because the highest-risk infants tend to be admitted to the lowest mortality NICUs. As shown in Figure 5, mortality measures that only adjust for birth weight tend to overestimate mortality rates among the units with low mortality. On average, providers in the lowest quintile of mortality according to the fully risk-adjusted filtered measures have nearly 1 percentage point higher mortality rates on the birth weight adjusted measure. Similarly, birth weight adjustment tends to understate mortality for the units identified as having the highest mortality rates with the fully risk-adjusted data. Thus, more complete risk adjustment as in the Network risk adjustment model contributes to more accurate comparisons across units.

Another dimension of the Network data on outcomes, which is both costly to collect and not commonly found in other administrative data systems, is the ability to follow infants across hospitals and track mortality after transfer from the first hospital. The correlation between the Network measure of 28-day filtered mortality (which tracks mortality after transfer) and a cruder version that only counts mortality in the initial hospital of admission was 0.91 in 1999,

suggesting that the benefits of mortality follow-up may be minor. However, the in-hospital version tends to understate mortality rates, particularly at units that transfer many infants out of the NICU in the first 28 days. Figure 6 reports the average bias in the mortality estimate (the difference between the filtered estimates of the crude measure and the Network measure) for hospitals that ranked in the bottom through the top quintile (20%) of the Network in terms of transfer rates. It is apparent that when mortality is only measured in the initial hospital, mortality at high-transfer units is biased downward by roughly one percentage point. Thus, the integrity of comparisons across these hospitals depends on the availability of follow-up mortality data.

#### *Comparisons Across Performance Measures*

In Table 3, we report the correlation in 1999 between filtered estimates of five alternative performance measures: risk-adjusted mortality rates at 3 days, 28 days, and overall, and risk-adjusted rates of nosocomial infection and severe intraventricular hemorrhage. The positive correlations among all of the measures indicate that NICUs doing well on any particular measure tend also to do well on the other measures. Overall mortality and 28-day mortality are strongly associated with each other (correlation=0.93), but less strongly associated with 3-day mortality (correlation=0.61, 0.70). Thus, any single mortality measure with at least 28-day follow-up may be sufficient to capture the dimensions of NICU quality that affect mortality at all times after admission. The overall and 28-day mortality rates are also more strongly correlated with rates of infection and intraventricular hemorrhage, further suggesting that these longer-term mortality rates are better indicators of differences in care across NICUs. This evidence is generally consistent with clinical experience, which suggests that care decisions made throughout the first month of life are the most important determinants of outcomes for VLBW infants. Finally, the

28-day mortality rate is only weakly related to the nosocomial infection rate (correlation=0.24). The fact that nosocomial infection carries different information than mortality suggests that this morbidity measure (in combination with mortality) would be particularly useful to doctors trying to understand the quality of care in their units.

## COMMENT

In the past, unreliability of conventional performance measures has justifiably limited the usefulness of actual performance measures in many practical applications, including the quality improvement efforts of the Vermont Oxford Network. For example, when such performance measures determine rewards or sanctions, providers are subjected unfairly to substantial risk of being punished or rewarded for outcomes beyond their control. Similarly, when such measures are used to identify best practice, virtually every style of medical practice is likely to have good results due to chance sooner or later. Finally, both provider and public confidence in quality indicators is likely to erode when provider performance appears to fluctuate wildly from one year to the next for no apparent reason. These types of problems have frustrated efforts to use performance measures in fields beyond health care as well. For example, recent legislation implementing a national school accountability program required changes after an analysis suggested that the accountability system was seriously compromised by its reliance on unreliable and unstable performance measures (Broder 2001; Kane, Staiger et al. 2001; Robelen 2001).

While conventional risk-adjusted outcome rates are problematic, our results suggest that more sophisticated approaches to estimating quality differences across providers can markedly improve the reliability and stability of performance measures. The improved measures allowed us to identify significant differences in patient outcomes between a large number of NICUs, and

allowed us to accurately identify units with persistently good outcomes for the purpose of benchmarking and identifying best practices. In addition, the improved stability of the filtered measures increased our ability to track performance over time and compare trends in each unit. Thus, the use of filtered measures in this application appears to overcome many of the most serious practical limitations of conventional performance measures.

While our analysis is limited to infants of very low birth weight admitted to hospitals with NICUs, the results are consistent with recent analyses of cardiac care (McClellan and Staiger 1999; McClellan and Staiger 2000; McClellan and Staiger 2000), HCUP measures (AHRQ 2001), and school test scores (Kane and Staiger 2001; Kane and Staiger 2002 (forthcoming)) in which the use of filtered measures led to similar improvements in reliability, stability and ability to prospectively identify best performers. All of these studies have found large differences in performance across units that were correlated across measures and over time, but that were largely obscured by random variations in conventional performance indicators.

There are costs of employing such a technique to characterize a hospital's performance. First, given the limited amount of information conveyed in any single year of data, a hospital may have to show several years of sustained improvements before the filtered measure registers a statistically significant improvement. This may complicate ongoing quality improvement efforts, which may have to gauge their effects in the short term on less reliable data. This is particularly true for small hospitals, which will take longer to muster enough evidence to warrant an improvement in their ranking. However, these costs must be weighed against the unreliability inherent in relying on conventional measures that have large random variations. Finally, the filtered estimates assume that there are patterns of variation and correlation across years that remain stable across providers and over time. However, such patterns may not apply to

providers that make dramatic changes in practice (for example, following a change in management) or in years with dramatic changes in financial incentives or technological innovation. In principle, such changes could be built into the statistical model describing changes in performance over time. That is, there is a tradeoff between improving performance measures for most health care providers and accounting for unusual circumstances at particular providers. Presumably, the best performance measures will try to balance both concerns.

The improved precision of the filtered quality estimates has other practical benefits. They can help clinicians focus on the areas and practices that most account for true differences in performance among providers, and thus hold the most potential for quality improvement. For example, our results show that, although mortality measures for very low birth weight infants are very correlated at different time intervals, provider differences continue to evolve all the way to 28 days after admission. Thus, important differences appear to exist throughout a NICU stay, and practices that reduce complications like intraventricular hemorrhage appear to be important in distinguishing better NICUs from worse ones. Moreover, because many outcome measures for very low birth weight infants are highly correlated, a single 28-day mortality measure may be sufficient to capture the dimensions of NICU quality that are related to mortality. However, some morbidity measures, such as nosocomial infection, appear to be useful indicators of non-mortality dimensions of NICU quality.

The statistical methods we use to construct filtered estimates are designed to address issues of reliability, but the extent to which they can address the second fundamental problem of bias depends on the quality of the underlying data and the extent to which patients differ across providers in ways that are difficult to measure. In general, the extent to which risk adjustment is needed to make fair comparisons is an empirical one. In the case of NICU comparisons for very

low birth weight infants, eliminating risk-adjusters from the data would materially affect hospital rankings. Thus, while more limited data would be less costly to collect, our results suggest that it would result in nontrivial bias against some of the better units.

Another central component of the Vermont Oxford Network's performance measures is that they are based on high quality data (Horbar and Leahy 1995). In particular, because the Network's emphasis is on confidential sharing for quality improvement rather than public reporting or high stakes accountability, its members have had less incentive to misrepresent their data or select patients in a manner that improves their measured performance. Clearly, concerns commonly raised about this type of gaming (Green and Wintfeld 1995; Hofer, Hayward et al. 1999) will be increasingly important as the stakes attached to individual performance measures rise.

Overall, our analysis suggests that significant and substantial quality differences exist across NICUs, and filtered quality measures are clearly superior to conventional measures in detecting such performance differences. As an indication of the practical value of these new measures, the Vermont Oxford Network is working to incorporate them into their quality reporting system over the next two years. But the potential application of this approach extends beyond the Network to many patient populations and types of providers. In any case where small numbers and many factors beyond providers' control generate instability and unreliability in conventional performance estimates, the greater reliability of filtered quality estimates may allow the detection of important provider-level differences in performance that previously have been viewed as practically important but impossible to measure reliably. Such instances seem more the norm than the exception in efforts to improve health care quality.



## REFERENCES

- AHCPR (1994). Version 1 HCUP Quality Indicators, Agency for Health Care Policy and Research.
- AHRQ (2001). Refinement of the HCUP Quality Indicators. Rockville, MD, Agency for Healthcare Research and Quality.
- Bentley, J. M. and D. B. Nash (1998). "How Pennsylvania hospitals have responded to publicly released reports on coronary artery bypass graft surgery " Jt Comm J Qual Improv **24**(1): 40-9.
- Berwick, D. M. and D. L. Wald (1990). "Hospital leaders' opinions of the HCFA mortality data." JAMA **263**(2): 247-9.
- Broder, D. S. (2001). Long Road to Reform: Negotiators Forge Education Legislation. Washington Post. Washington, DC: 1.
- Burgess, J. F., Jr., C. L. Christiansen, et al. (2000). "Medical profiling: improving standards and risk adjustments using hierarchical models." J Health Econ **19**(3): 291-309.
- Chamberlain, G. (1980). "Analysis of Covariance with Qualitative Data." Review of Economic Studies **48**: 225-238.
- Davies, H. T. and J. Lampel (1998). "Trust in performance indicators? ." Qual Health Care **7**(3): 159-62.
- Davies, H. T. and M. N. Marshall (1999). "Public disclosure of performance data: does the public get what the public wants?" Lancet **353**(9165): 1639-40.
- Epstein, A. M. (1998). "Rolling down the runway: the challenges ahead for quality report cards." JAMA **279**(21): 1691-6.
- Green, J. and N. Wintfeld (1995). "Report cards on cardiac surgeons. Assessing New York State's approach." N Engl J Med **332**(18): 1229-32.
- Greene (1997). Econometric Analysis. New Jersey, Prentice Hall.
- Guyer, B., M. MacDorman, et al. (1998). "Annual Summary of Vital Statistics - 1997." Pediatrics **102**(6).
- Hannan, E. L., H. Kilburn, Jr., et al. (1994). "Improving the outcomes of coronary artery bypass surgery in New York State." JAMA **271**(10): 761-6.
- Hannan, E. L., M. J. Racz, et al. (1997). "Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough?" Health Serv Res **31**(6): 659-78.
- Harris, C. (2001). "England introduces star system for hospital trusts." BMJ **323**(709): 709.
- Hibbard, J. H. and J. J. Jewett (1997). "Will quality report cards help consumers?" Health Aff (Millwood) **16**(3): 218-28.
- Hibbard, J. H., J. J. Jewett, et al. (1997). "Choosing a health plan: do large employers use the data?" Health Aff (Millwood) **16**(6): 172-80.
- Hofer, T. P. and R. A. Hayward (1995). "Can early re-admission rates accurately detect poor-quality hospitals?" Med Care **33**(3): 234-45.
- Hofer, T. P. and R. A. Hayward (1996). "Identifying poor-quality hospitals. Can hospital mortality rates detect quality problems for medical diagnoses?" Med Care **34**(8): 737-53.
- Hofer, T. P., R. A. Hayward, et al. (1999). "The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease " JAMA **281**(22): 2098-105.
- Horbar, J. D. (1999). "The Vermont Oxford Network: evidence-based quality improvement for neonatology." Pediatrics **103**(1 Suppl E): 350-9.

- Horbar, J. D. and e. al. (2001). "Collaborative Quality Improvement for Neonatal Intensive Care." Pediatrics **107**(1): 14-22.
- Horbar, J. D. and K. A. Leahy (1995). "for the investigators of the Vermont-Oxford Trials Network: An Assessment of Data Quality in the Vermont-Oxford Trials Network Database." Controlled Clinical Trials **16**: 51-61.
- Iezzoni, L. I., A. S. Ash, et al. (1995). "Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes." Ann Intern Med **123**(10): 763-70.
- Iezzoni, L. I., A. S. Ash, et al. (1996). "Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method." Am J Public Health **86**(10): 1379-87.
- Iezzoni, L. I., A. S. Ash, et al. (1997). "Differences in procedure use, in-hospital mortality, and illness severity by gender for acute myocardial infarction patients: are answers affected by data source and severity measure?" Med Care **35**(2): 158-71.
- Iezzoni, L. I., M. Schwartz, et al. (1995). "Using severity-adjusted stroke mortality rates to judge hospitals." Int J Qual Health Care **7**(2): 81-94.
- Iezzoni, L. I., M. Schwartz, et al. (1996). "Severity measurement methods and judging hospital death rates for pneumonia." Med Care **34**(1): 11-28.
- Iezzoni, L. I., M. Schwartz, et al. (1996). "Predicting in-hospital mortality for stroke patients: results differ across severity-measurement methods." Med Decis Making **16**(4): 348-56.
- Iezzoni, L. I., M. Schwartz, et al. (1996). "Using severity measures to predict the likelihood of death for pneumonia inpatients." J Gen Intern Med **11**(1): 23-31.
- Kane, T. J. and D. O. Staiger (2001). Improving School Accountability Measures.
- Kane, T. J. and D. O. Staiger (2002 (forthcoming)). Volatility in School Test Scores: Implications for Test-Based Accountability Systems. Brookings Papers on Education Policy, 2002. D. Ravitch. Washington, DC, Brookings Institution.
- Kane, T. J., D. O. Staiger, et al. (2001). Assessing the Definition of "Adequate Yearly Progress" in the House and Senate Education Bills. UCLA, School of Public Policy and Social Research. Los Angeles, CA.
- Krumholz, H. M., J. Chen, et al. (1999). "Comparing AMI mortality among hospitals in patients 65 years of age and older: evaluating methods of risk adjustment." Circulation **99**(23): 2986-92.
- Localio, A. R., B. H. Hamory, et al. (1997). "The public release of hospital and physician mortality data in Pennsylvania. A case study." Med Care **35**(3): 272-286.
- Luft, H., P. Romano, et al. (1996). Second Report of the California Hospital Outcomes Project: Acute Myocardial Infarction. O. o. S. H. P. a. Development, Office of Statewide Health Planning and Development.
- Marshall, M. N., P. G. Shekelle, et al. (2000). "The public release of performance data: what do we expect to gain? A review of the evidence." Jama **283**(14): 1866-74.
- McClellan, M. and D. Staiger (1999). The quality of health care providers.
- McClellan, M. and D. Staiger (2000). Comparing Hospital Quality at For-Profit and Not-for-Profit Hospitals. The Changing Hospital Industry. D. Cutler. Chicago, IL, University of Chicago Press: 93-112.
- McClellan, M. and D. Staiger (2000). Comparing the Quality of Health Care Providers. Frontiers in Health Policy. A. M. Garber. Cambridge, MA, The MIT Press. **3**: 113-136.
- Mennemeyer, S. T., M. A. Morrissey, et al. (1997). "Death and reputation: how consumers acted upon HCFA mortality information." Inquiry **34**(2): 117-28.
- Morris, C. N. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." JASA **78**(381): 47-55.

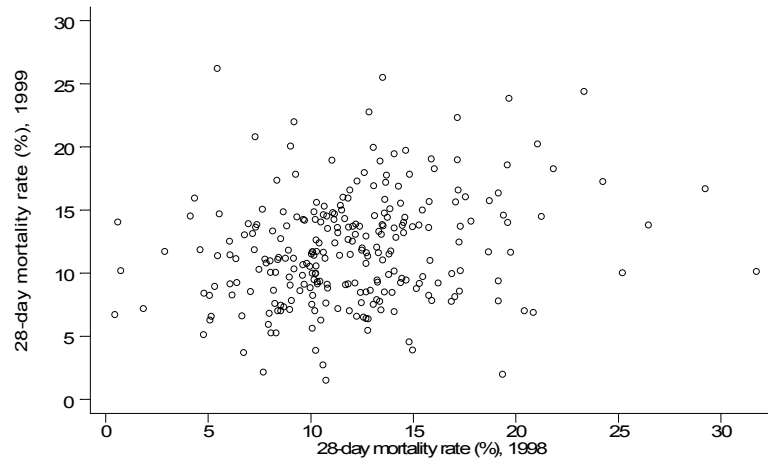
- Normand, S., M. Glickman, et al. (1997). "Statistical methods for profiling providers of medical care: Issues and applications." JASA **92**(439): 803-814.
- O'Connor, G. T., S. K. Plume, et al. (1996). "A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. The Northern New England Cardiovascular Disease Study Group." JAMA **275**(11): 841-6.
- O'Hagan, A. (1994). Bayesian Inference. Kendall's Advanced Theory of Statistics. G. P. e. al. New York, Halstead Press. **2B**.
- Papile, L., J. Burstein, et al. (1978). "Incidence and Evolution of Subependymal and intraventricular Hemorrhage: a study of infants with birthweights less than 1500 grams." Journal of Pediatrics **92**: 529-534.
- Park, R. E., R. H. Brook, et al. (1990). "Explaining variations in hospital death rates. Randomness, severity of illness, quality of care." JAMA **264**(4): 484-90.
- Parry, G. J., C. R. Gould, et al. (1998). "Annual league tables of mortality in neonatal intensive care units: longitudinal study. International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group." BMJ **316**(7149): 1931-5.
- Pine, M., M. Norusis, et al. (1997). "Predictions of hospital mortality rates: a comparison of data sources." Ann Intern Med **126**(5): 347-54.
- Robelen, E. (2001). Bush warns against 'undoable' ESEA progress standard. Education Week. Washington, DC.
- Schneider, E. C. and A. M. Epstein (1996). "Influence of cardiac-surgery performance reports on referral practices and access to care. A survey of cardiovascular specialists." N Engl J Med **335**(4): 251-6.
- Schneider, E. C. and A. M. Epstein (1998). "Use of public performance reports: a survey of patients undergoing cardiac surgery." JAMA **279**(20): 1638-42.
- Thomas, J., J. Holloway, et al. (1993). "Validating risk-adjusted mortality as an indicator for quality of care." Inquiry **30**(1): 6-22.
- Thomas, J. W. (1996). "Does risk-adjusted readmission rate provide valid information on hospital quality?" Inquiry **33**(3): 258-70.
- Thomas, J. W. and T. P. Hofer (1998). "Research evidence on the validity of risk-adjusted mortality rate as a measure of hospital quality of care [published erratum appears in Med Care Res Rev 1999 Mar;56(1):118]." Med Care Res Rev **55**(4): 371-404.
- Thomas, J. W. and T. P. Hofer (1999). "Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care." Med Care **37**(1): 83-92.

**FIGURES AND TABLES**

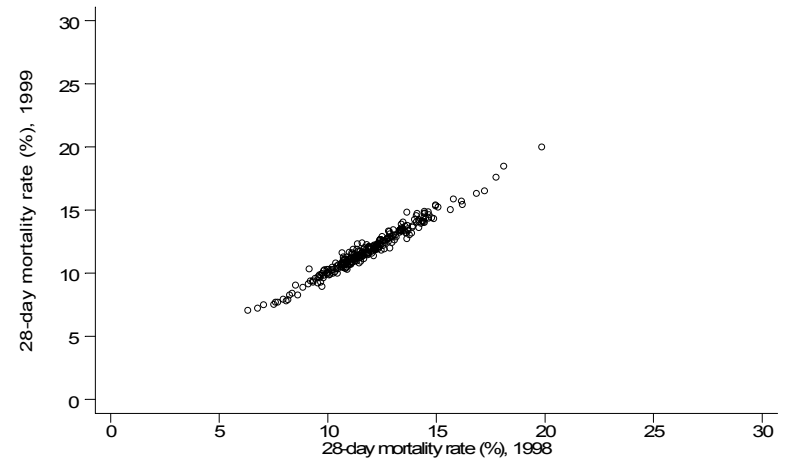


**Figure 2. Conventional (Left) and Filtered (Right) Estimates of Risk-Adjusted 28-day Mortality Rates, 1998 Estimates Plotted Against 1999 Estimates**

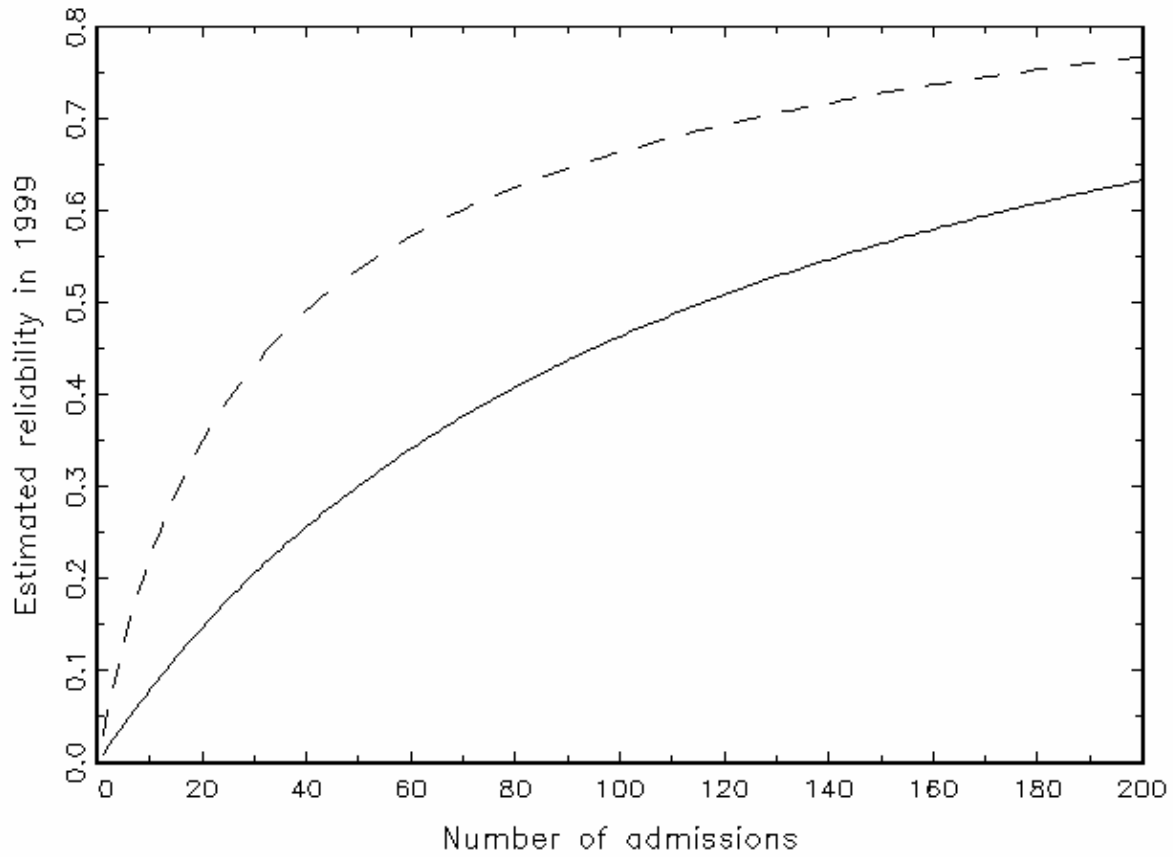
*A. Conventional Estimates*



*B. Filtered Estimates*



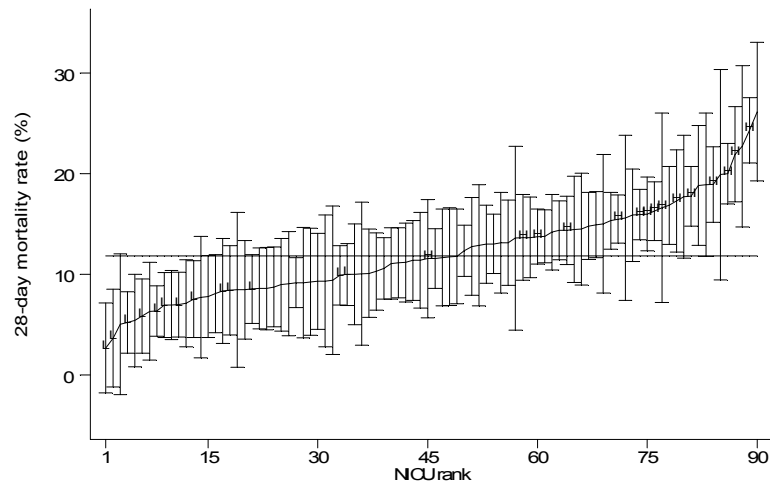
**Figure 3. Reliability of Filtered (Top Line) and Conventional (Bottom Line) Estimates of 28-day Mortality Rates**



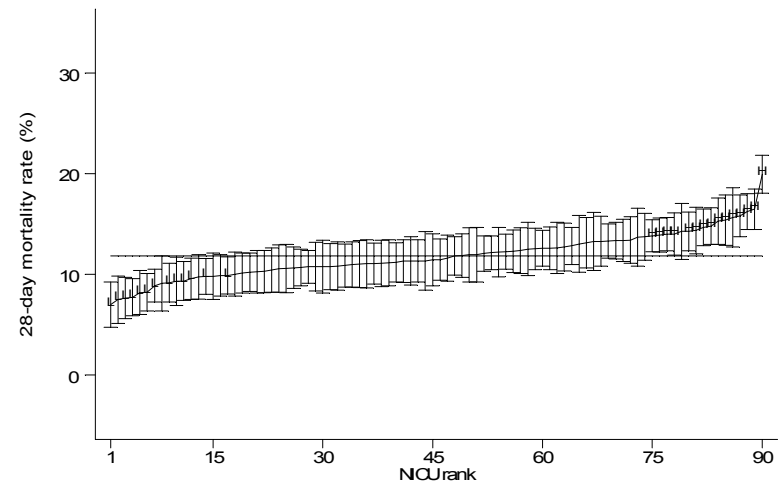
**Figure 4. Rankings with 1.4SE Intervals Based on Conventional (Left) and Filtered (Right) Estimates of Risk-Adjusted 28-day Mortality for 1999**

“H” indicates high mortality based on filtered estimates, “L” indicates low mortality based on filtered estimates.

*A. Conventional Estimates*

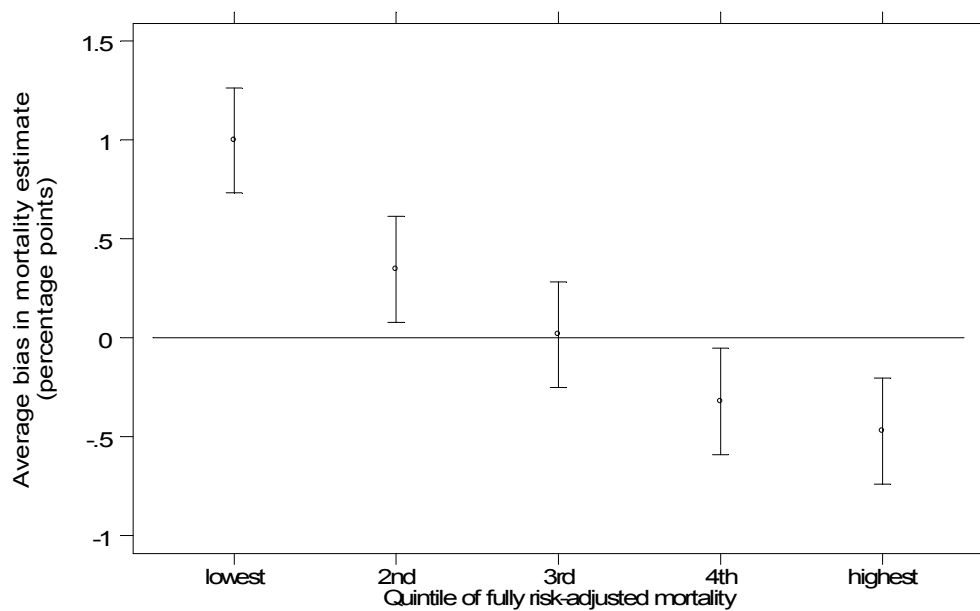


*B. Filtered Estimates*



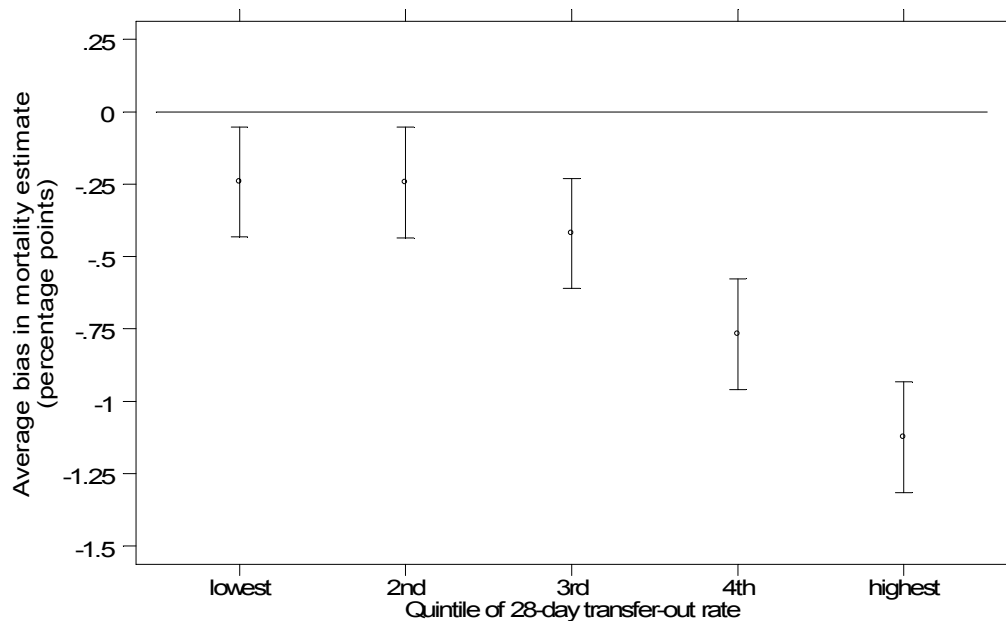
### Figure 5. Bias in Filtered Estimates of Risk-Adjusted 28-day Mortality if only Risk-Adjust for Birth weight.

Each bar gives the average difference between birth weight adjusted and fully risk-adjusted filtered estimates of 28-day mortality, by mortality quintile based on the fully risk-adjusted measures. Estimates are significantly different across quintiles ( $p < .001$ ).



### Figure 6. Bias in Filtered Estimates of Risk-Adjusted 28-day Mortality if only Count Mortality Occurring at the Initial Hospital of Admission.

Each bar gives the average difference between risk-adjusted filtered estimates of initial hospital versus all 28-day mortality, by quintile according each hospital's rate of transfer out within 28 days of birth. Estimates are significantly different across quintiles ( $p < .001$ ).



**Table 1. Estimates of the Variation in Outcome Rates across NICUs in 1999, and the Correlation with Outcome Rates from the Prior Year.**

<i>Risk-Adjusted Outcome Measures</i>	<b>Mean Outcome (%)</b>	<b>Conventional Estimates</b>		<b>Estimates Corrected for Estimation Error in Outcome Measures</b>	
		<b>Standard Deviation (%)</b>	<b>Correlation with prior year</b>	<b>Standard Deviation (%)</b>	<b>Correlation with prior year</b>
3-day Mortality Rate	6.1 (0.2)	3.5 (0.3)	0.32 (0.08)	1.9 (0.2)	0.85 (0.06)
28-day Mortality Rate	11.8 (0.3)	5.0 (0.3)	0.27 (0.06)	2.5 (0.2)	0.89 (0.06)
Overall Mortality Rate	14.3 (0.3)	5.6 (0.3)	0.34 (0.06)	3.0 (0.2)	0.91 (0.05)
Nosocomial Infection Rate	26.1 (0.6)	10.8 (0.5)	0.55 (0.05)	8.4 (0.4)	0.87 (0.03)
Severe Intra-Ventricular Hemorrhage Rate	17.7 (0.4)	7.1 (0.8)	0.38 (0.08)	3.2 (0.3)	1.0 (0.05)

Note: Standard errors of the estimates are in parentheses.

**Table 2. Using Filtered Estimates of Best Performers in 1997 to Identify Mortality Differences in 1999**

	Rankings based on 1997 filtered estimates of risk-adjusted 28-day mortality		
	Top 10% NICUs (low mortality)	Middle 80% NICUs (average mortality)	Bottom 10% NICUs (high mortality)
Average risk-adjusted 28-day mortality rate in 1999 (standard error)	8.0 (0.9)	11.8 (0.3)	13.5 (1.1)
Filtered estimate of risk-adjusted 28-day mortality rate in 1999, based on data from 1994-1997	9.2	11.8	14.5
Conventional estimate of risk-adjusted 28-day mortality rate in 1997	7.2	11.9	20.9

Note: Based on 222 NICUs with data for both 1997 and 1999.

**Table 3. Correlation Between Alternate Risk-Adjusted Filtered Performance Measures in 1999**

	28-day mortality rate	Overall mortality rate	Nosocomial Infection rate	Severe Intra-ventricular hemorrhage rate
3-day mortality rate	0.70	0.61	0.13	0.48
28-day mortality rate		0.93	0.24	0.75
Overall mortality rate			0.30	0.75
Nosocomial Infection rate				0.25

N=297 NICUs with data in 1999.

## TECHNICAL APPENDIX

Deriving filtered estimates proceeds in 3 steps. First, we calculate conventional quality indicators for each hospital and for as many years as data are available. Second, we estimate the variance of the estimation error for the observed quality indicators (the “noise” variance) and the variance of the true quality differences across hospitals (the “signal” variance). Finally, we form predictions of true quality differences across hospitals in each year based on a weighted average of the data from all years, using optimal weights derived from our estimates of the signal and noise variance. We refer to these as “filtered” estimates, since these estimates are attempting to filter out the estimation error in the conventional quality indicators (and because our method is closely related to the idea of filtering in statistics). We provide an overview of each of these steps below. For more detail (including how to generalize this approach to allow for multiple quality indicators per year) see McClellan and Staiger (McClellan and Staiger 1999).

Step 1: Patient-level Regressions . Suppose we have a sample of patients admitted to  $N$  hospitals over  $T$  years. For each patient (indexed by  $I$ ), we have an outcome measure of interest,  $Y$  (e.g. mortality within 28 days), and a vector of patient covariates,  $X$  (e.g. birth weight, agar score and other risk factors). For each year, we construct a conventional quality indicator based on the (risk-adjusted) average of the outcome measure ( $Y$ ) in each hospital, i.e. from a patient-level regression of the form:

$$(1) \quad Y_{jti} = \delta_{jt} + X_{jti}\beta + u_{jti}$$

Estimation of equation (1) by standard regression methods yields unbiased estimates of the hospital-specific intercepts ( $\hat{\delta}_{jt}$ ) for hospital  $j$  in year  $t$ , along with the standard errors on these estimates ( $SE_{\hat{\delta}_{jt}}$ ). These hospital-specific intercepts are what we refer to as conventional risk-adjusted outcome measures, i.e. they are the conventional estimates of risk-adjusted outcome rates for each hospital in each year.

We use linear regression rather than logistic regression for three reasons. First, previous work (McClellan and Staiger 1999; McClellan and Staiger 2000; AHRQ 2001) on an adult population has successfully used OLS regression. Second, estimating logistic models with hospital fixed effects cannot be done with standard methods, and requires computationally intensive conditional maximum likelihood methods (Chamberlain 1980). Finally, the widely-used approach of estimating a logistic model without hospital effects, and then using the resulting predictions as estimates of the expected outcome, yields inconsistent parameter estimates and inconsistent predictions when hospital quality is correlated with patient case mix, as is likely to be the case in our application. In contrast, it is easy to incorporate hospital fixed effects into OLS regression, the resulting estimates are consistent, and the hospital fixed effects provide direct estimates of the probability of the outcome at each hospital, holding constant all observed patient characteristics. One limitation of the OLS approach is that it assumes that a linear probability model is the correct functional form. However, as in earlier work with heart attack outcomes, we specify a fairly flexible functional form so that the linear probability model will provide a flexible approximation to any more general probability model. A second limitation of the OLS approach is that it yields inefficient (but unbiased) estimates because the errors of a linear probability model are necessarily heteroskedastic. Given the large sample sizes

for these regressions (approximately 20,000 infants per year) efficiency is not likely to be an important concern.

Step 2: Estimating Noise and Signal Variance. Conventional quality indicators ( $\hat{\delta}_{jt}$ ) are based on a sample of patients, and therefore are noisy estimates of the true quality levels ( $\delta_{jt}$ ) that we really wish to measure. In any given year, even a high quality hospital may have poor patient outcomes because of chance events. This is particularly true when a quality indicator is based on only a handful of patients, so that the outcome of one or two patients will materially influence that hospital's estimate. Thus, we can think of the observed vector of quality indicators ( $\hat{\delta}_j = \{\hat{\delta}_{j1}, \dots, \hat{\delta}_{jT}\}$ ) as estimates of the true quality indicators ( $\delta_j = \{\delta_{j1}, \dots, \delta_{jT}\}$ ) that are of interest:

$$(2) \quad \hat{\delta}_j = \delta_j + \varepsilon_j$$

Where  $\varepsilon_j$  is the estimation error (which will tend to be larger for hospitals with smaller samples of patients). Thus, the conventional quality indicators are composed of signal ( $\delta_j$ ) and noise ( $\varepsilon_j$ ) components.

As discussed above, estimates of the signal and noise variance are necessary to construct filtered estimates of hospital quality. We estimate each of these as follows.

The noise variance ( $\text{Var}(\varepsilon_j)$ ) is related to the precision of the estimated quality indicators. Therefore, the standard error for each year's quality indicator ( $SE_{\hat{\delta}_{jt}}$ ) can be used to estimate the noise variance. In particular, we estimate the noise variance in each year with  $\text{Var}(\varepsilon_{jt}) = (SE_{\hat{\delta}_{jt}})^2$ .

We must also estimate any covariance in the estimation error across years. Since each year's

measure is based on a separate sample of patients, the covariance in the estimation error from one year to the next is zero.

The signal variance ( $\text{Var}(\delta_j)$ , a  $T \times T$  matrix) captures how much true quality varies across hospitals as well as how correlated the true quality differences are over time. To estimate the signal variance in each year, we note that the variance of the estimated quality indicators ( $\text{Var}(\hat{\delta}_{jt})$ ) is simply the sum of the signal and noise variance ( $\text{Var}(\delta_{jt}) + \text{Var}(\varepsilon_{jt})$ ). Thus, to estimate the signal variance we calculate the total variance across hospitals (the sample variance of  $\hat{\delta}_{jt}$ ) and subtract the amount expected to be due to noise variance (the sample average of  $\text{Var}(\varepsilon_{jt})$ ). We must also estimate any covariance in the signal across years. Since the estimation error is independent from year to year, the covariance in the signal can be estimated directly using the sample covariance of the conventional quality indicators (e.g.  $\text{Cov}(\hat{\delta}_t, \hat{\delta}_s)$ ).

Finally, we summarize the signal variance in each year along with the correlation in the signal between years with a simple two-parameter first-order auto-regressive model. This model assumes that the signal variance is stable over time, and that the correlation in the quality signal between any two years declines geometrically in proportion to the length of time between the years. We estimate the parameters of the auto-regressive model using General Method of Moments (Greene 1997). This method chooses the parameters so as to fit the unconstrained estimates of the signal variance as closely as possible. For details of how this model is estimated and its empirical performance see earlier work by McClellan and Staiger (McClellan and Staiger 1999).

Step 3. Constructing Filtered Estimates. Our problem is how to use the entire history of observed quality indicators for each hospital ( $\hat{\delta}_j$ ) to best predict the true quality for any given hospital and year ( $\delta_{jt}$ ). For simplicity, suppose that we have subtracted the average outcome in each year off of both  $\hat{\delta}_j$  and  $\delta_{jt}$  so that we are trying to predict the difference from the average hospital in each hospital's quality. Conceptually, we can think of our problem as similar to that of minimizing the prediction errors ( $v$ ) in a regression of the form:

$$(3) \quad \delta_{jt} = \hat{\delta}_j \beta + v_{jt}$$

Where the regression coefficients  $\beta$  are the optimal weights to place on each year's quality indicator in order to best predict the true quality level in year  $t$ . One would expect these weights to vary by hospital and year, since the precision of the estimated quality indicators varies by hospital and year. Note that if  $\beta$  were zero, this would lead to a prediction that the hospital's quality was equal to the average hospital (e.g., the prediction would place all of the weight on the overall average and none on the hospital's own quality estimates).

We cannot run a simple regression to estimate the weights in equation 3 since the true quality level is unobserved. The conventional approach simply uses the estimated quality indicators in year  $t$  as the best estimates of the true quality level in year  $t$ . This is equivalent to assuming that each quality indicator has no error and, by itself, is the best predictor of the true quality level – i.e. that the coefficient on  $\hat{\delta}_{jt}$  is equal to 1 in equation (3), and all other coefficients are zero.

But there are two problems with this approach. First, since the quality indicator in year  $t$  is estimated with error, we can improve the mean squared error of the prediction by attenuating

its coefficient towards zero. Moreover, this attenuation should be greater for hospitals in which the quality indicators are not precisely estimated. This is the basic idea behind Bayesian shrinkage estimators (Morris 1983): the observed variation in quality indicators will tend to overstate the amount of actual variation across hospitals, so by pulling all the estimates (especially the more imprecise estimates) back toward the mean we can improve prediction accuracy. The second problem with the conventional method is that it does not use any of the information available in other quality indicators. If the true quality level in other years is correlated with this year's quality, then using the information in estimates of these other quality indicators can further improve prediction accuracy.

Filtered estimates are a simple method for creating predictions of  $\delta_{jt}$ , based on equation 3, which incorporate both the shrinkage idea and information from all available quality indicators. In particular, note that estimating the optimal weights in equation 3 is analogous to estimating regression coefficients. The standard formula for estimating the regression coefficients in equation 3 is:

$$(4) \quad \hat{\beta} = [Var(\hat{\delta}_j)]^{-1} Cov(\hat{\delta}_j, \delta_{jt}) = [Var(\delta_j) + Var(\varepsilon_j)]^{-1} Cov(\delta_j, \delta_{jt})$$

Thus, the optimal weights depend only on the ratio of signal variance (and covariance) to signal plus noise variance. We estimate these optimal weights using estimates of the signal and noise variance from step 2, and then create filtered estimates as the predicted value from equation 3, e.g. as  $\hat{\delta}_j \hat{\beta}$ . Standard errors for these predictions are calculated in an analogous method, using conventional formulas for the standard error of a regression prediction and using estimates of all necessary moment matrices from step 2. Similarly, estimates of reliability for the

filtered measures are constructed using conventional formulas for the R-squared from the regression equation (3), and using estimates of all necessary moment matrices from step 2.

To summarize, our method proceeds in 3 stages. First, we calculate a vector of conventional quality indicators for each hospital and for as many years as data are available. Second, we estimate the variance of the estimation error for the observed quality indicators ( $\text{Var}(\varepsilon)$ ) and the variance of true quality across hospitals ( $\text{Var}(\delta)$ ) as discussed above. Finally, we use these variance estimates to form predictions of true quality at each hospital based on equation (3), using optimal weights derived from equation (4). We refer to these predictions as filtered estimates.