2001

Benchmarking patient outcomes

Ellen B. Rudy
Joseph F. Lucke, *State University of New York at Buffalo*
Gayle R. Whitman
Lynda J. Davidson

Benchmarking Patient Outcomes

Ellen B. Rudy, Joseph F. Lucke, Gayle R. Whitman, Lynda J. Davidson

Purpose: *To examine the usefulness of three types of benchmarking for interpreting patient outcome data.*

Design: *This study was part of a multiyear, multihospital longitudinal survey of 10 patient outcomes. The patient outcome used for this methodologic presentation was central line infections (CLI). The sample included eight hospitals in an integrated healthcare system, with a range in size from 144 to 861 beds. The unit of analysis for CLI was the number of line days, with the CLI rate defined as the number of infections per 1,000 patient-line days per month.*

Methods: *Data on each outcome were collected at the unit level according to standardized protocols. Results were submitted via standardized electronic forms to a central data management center. Data for this presentation were analyzed using a Bayesian hierarchical Poisson model. Results are presented for each hospital and the system as a whole.*

Findings: *In comparison to published benchmarks, hospital performances were mixed with regard to CLI. Five of the 8 hospitals exceeded 2.2 infections per 1,000 patient-line days. When benchmarks were established for each hospital using 95% credible intervals, hospitals did reasonably well with only isolated months reaching or going beyond the benchmark limits. When the entire system was used to establish benchmarks with the 95% credible intervals, the hospitals that reached or exceeded the benchmark limits remained the same, but some hospitals had CLI rates more frequently in the upper 50% of the benchmarking limits.*

Conclusions: *Benchmarking of quality indicators can be accomplished in a variety of ways as a means to quantify patient care and identify areas needing attention and improvement. Hospital-specific and system-wide benchmarks provide relevant feedback for improving performance at individual hospitals.*

JOURNAL OF NURSING SCHOLARSHIP, 2001; 33:2, 185-189.

[Key words: benchmarking, central line infections]

* * *

Across the United States managing the costs and quality of patient outcomes continues to be a driving force in the healthcare industry. Although the primary focus of managed care networks has been on cost control and competitive strategies, the concern about quality has also increased (Byrne, Schreiner, Rizk & Sokolowski, 1998; Epstein, 1998; Office of Technology and Assessment, 1995). Indicators of quality of care include specific patient outcomes, particularly mortality (Baggs, Ryan, Phelps, Richeson, & Johnson, 1992; Fink, Yano, & Brook, 1989; Knaus, Draper, Wagner, & Zimmerman, 1986; Mitchell, Armstrong, Simpson, & Lentz, 1989; Paneth et al., 1982; Scott, Flood, & Ewy, 1979; Shortell et al., 1994). The public, however, is no longer satisfied with such a narrow definition of quality

Ellen B. Rudy, RN, PhD, FAAN, *Eta*, Dean and Professor; **Joseph F. Lucke**, PhD, *Eta*, Associate Professor; **Gayle R. Whitman**, RN, PhD, FAAN, *Eta*, Associate Professor; **Lynda J. Davidson**, RN, PhD, *Eta*, Associate Dean and Assistant Professor; all at the University of Pittsburgh School of Nursing, Pittsburgh, PA. The authors acknowledge their colleagues from the UPMC Health Care System for their support in this work: Gail A. Wolf, RN, DnSc, FAAN, Chief Nursing Officer, UPMC Health System; Joyce Lewis, RN, MSN, Director of Nursing, UPMC Beaver; Mary Anne Cozza, RN, MSN, Vice President, Patient Care Services, UPMC Braddock; Cheryl Como, RN, MSN, Vice President, UPMC McKeesport; Sandra McCarthy, RN, MSOL, Vice President, UPMC Passavant; Andrea Schmid, RN, MBA, MSN, Vice President, UPMC Presbyterian - Montefiore; Tamra Merryman, RN, MSN, CHE, Vice President, UPMC Shadyside; Beverly Haines, RN, MNEd, Vice President, UPMC Southside and Katherine Vidakovich, RN, MSEd, Vice President, UPMC St. Margaret. Correspondence to Dr. Rudy, University of Pittsburgh School of Nursing, 350 Victoria Building, Pittsburgh, PA 15261. E-mail: erudy@pitt.edu
Accepted for publication August 21, 2000.

care, and consumers are putting increased pressure on health care organizations to provide data on multiple indicators of quality.

The general trend has been to examine patient outcomes as measures of quality to quantify outcomes and to use large databases to answer questions about quality (Aiken, Sloane, Lake, Sochalski, & Weber, 1999; Lichtig, Knauf, & Milholland, 1999). Difficulties in this process include: (a) few consistent definitions of patient outcome variables, (b) no standard for frequency of measuring each outcome (e.g., daily, weekly, monthly), and (c) few published benchmarks that allow hospital staff or network system administrators to determine whether they are meeting a quality indicator.

The purpose of this methodologic study was to present three methods that can be used to establish an expected or "normal" level of quality, or a "benchmark" that can be used to judge high quality care. The methods used to illustrate the usefulness of establishing benchmarks for interpreting patient outcome data were (a) an examination of the literature from prior studies, (b) a statistical method that used data collected on the specific patient outcome within each individual hospital so that the benchmark obtained was hospital-specific, and (c) a statistical method to aggregate all the data among hospitals to determine a system-wide benchmark. Central line infections were chosen as the patient outcome to demonstrate the benchmarking methods. CLI are a costly complication to patient care and occur in highly vulnerable patients, making them useful markers of quality.

Background

Benchmarking is a term that comes from surveyors who marked posts, rocks, and other objects to indicate a starting point for determining altitudes. It still refers to a starting point or a point of reference. Although the term "benchmarking" has become commonplace in the United States, people in other countries may not be familiar with this term and might find the term "standard setting" more applicable. Within the U.S. healthcare industry, benchmarking has come to mean: The level of cost for a specific product. How much does it cost, for example, to do coronary artery bypass surgery? The national or regional benchmark for the cost of this surgery becomes the cheapest that it can be done within a certain level of mortality and morbidity. Benchmarking is only as good as the accuracy of the reporting from healthcare facilities and the willingness of personnel in different facilities to share this information. Because third-party payers have such concerns about costs, they often report the costs of various medical procedures across institutions.

The concern for quality care now extends well beyond the original focus on cost. Both health care leaders and the public recognize that delays in recovery, evidence of complications, and changes in functional status are outcomes that also indicate quality of care and deserve as much attention as do the benchmarks of cost.

The three examples in this paper illustrate methods other than cost that can be used to establish benchmarks. The first presented is the method based on a review of prior studies reported in the literature, the second method results in hospital-specific benchmarks, and the third is a method for system-wide benchmarks.

Method 1: Review of Literature

Three large studies illustrate the mechanism for external benchmarking. In a national prevalence study of 72 hospitals in Germany, catheter-associated bloodstream infections represented 8.3% of all nosocomial infections. Among 55,400 central venous catheter days in 14,988 ICU patients, 2.2 catheter-associated primary bloodstream infections per 1,000 catheter days were reported (95% confidence interval 1.8-2.6) (Gastmeier, Weist, & Ruden, 1999).

In a study over 28 months in a Veterans' Administration Medical Center, 300 catheters were inserted in 204 patients. Bacteremia occurred in patients in 2.7% of catheter insertions, insertion-site infection developed in 1.3%, and catheter colonization developed in 12%. Relationship of infection to number of catheter days was not reported (Goetz, Wagener, Miller, & Muder, 1998).

New England investigators, comparing central-line infections with the number of catheter line-days in place, showed 3.98 CLI per 1,000 catheter days (95% confidence interval, 2.06 to 6.96) for the total cohort of patients, but a higher rate of 4.2 per 1,000 catheter days (95% confidence interval, 1.81 to 8.29) for ICU patients. The New England study showed only 400 patients with 3,014 catheter-line days but the rate of CLI was higher than rates in the study from Germany (Gowardman et al., 1998). Other studies have been reported on the use of antiseptic-impregnated central line catheters compared to use of standard catheters as a means of decreasing infection rates. However, those studies include little comparable data that can be used for clinical benchmarking because they do not show the number of catheter-line days.

As with all clinical research, researchers in these studies had problems with consistency in definitions, how data were collected and analyzed, and applicability to different settings and different patient populations. A reasonable conclusion based on these studies is that CLI occur at a rate between 2.2 to 4.2 per 1,000 catheter days, and higher rates can be expected to occur in ICU patient populations.

Method 2: Hospital-Specific Benchmarks

Data for this study were collected over 1 year as part of a multiyear, multihospital longitudinal survey of 10 patient outcomes to establish benchmark goals. The eight hospitals in this report included two urban tertiary-care hospitals, five urban community hospitals, and one rural community hospital, ranging in size from 144 to 861 beds. The process of establishing definitions and assuring reliability and validity of data are reported elsewhere (Whitman, Davidson, Rudy, & Wolf, 2001). These data were collected for each month in each hospital from January to December 1998. All data were submitted to a central data management center via standardized electronic forms, specific to each outcome variable. Upon receipt, research team personnel further scrutinized the data for errors. Data were stored via computer in Microsoft Excel. Prior to statistical analysis, the data were again reviewed for distributional anomalies by exploratory data analytic methods in S-PLUS 2000 (1999).

The unit of analysis was 1,000 patient-line days per month. The monthly 1,000 patient-line days aggregated over all eight hospitals ranged from .064 to 2.056 with a median of .37 and half the data falling between .215 and .537. The annual number of 1,000 patient-line days for each hospital ranged from 1.561 to 21.140, with a total of 46.701 (**Table 1**). The outcome variable was the number of central-line infections per 1,000 line days per month. Standard surveillance protocols and nosocomial infection site definitions from the National Nosocomial Infection Surveillance (NNIS) System of the Center for Disease Control and Prevention (1999) were used at all sites.

Within each hospital, each patient was assumed to have a hospital-specific risk of CLI based solely on the number of days exposed to the central line (line-days). This hospital-specific risk was assumed to be identical for each patient, independent from one patient to another and constant over the entire year of observation. Thus, the number of CLI per month for each hospital was strictly a function of the number of patient-line days.

The data were assumed to follow a Poisson distribution, the canonical distribution for count data (Bernardo & Smith, 1994). Missing data were assumed to be missing at random: the process generating missing data was either completely random or a random function of the observed data (Little & Rubin, 1987). The hospital-specific rates of CLI were estimated by standard Bayesian methods for the Poisson distribution (Bernardo & Smith, 1994), using WinBUGS 1.2 (Spiegelhalter, Thomas, Best, & Gilks, 1999).

Given the estimated hospital-specific rates, a predicted median (50^{th} percentile) number of CLI (based on the number of 1,000 patient-line days for that month) together with a 95% credible interval (2.5 and 97.5 percentiles) were calculated for each month for each hospital and plotted over 1 year. Credible intervals are not confidence intervals. Credible intervals are established after the data are collected and indicate the range in which 95% of the data can be expected to fall (Bernardo & Smith, 1994). In contrast, a 95% confidence interval is established before collecting any data and it is but one of a hypothetically infinite sequence of intervals covering the data 95% of the time.

Once the 95% credible intervals are established, any observed points exceeding the upper limits of the intervals are flagged as outside the acceptable benchmark. For a given rate, a larger number of 1,000 patient-line days will yield larger estimated percentiles and a larger width of the credible interval. A higher rate of CLI will yield higher estimated percentiles and a wider credible interval.

Results

The monthly frequencies of CLI ranged from 0 to 15 with a median of 1 and half the frequencies between 0 and 2. The monthly rates of CLI per 1,000 patient-line days ranged from 0 to 15.5 with a median of 2.79 and half the rates between 0 and 4.28.

Table 1 shows the hospital-specific rates of CLI and their respective standard errors. The rates ranged from a low of 1.62 (Hospital H) to a high of 4.87 (Hospital E) CLI per 1,000 patient-line days.

Table 1. Rates of Central Line Infections per 1,000 Patient-Line Days

Hospital	1,000 Patient-line days	Hospital specific mean (std. error)	System based mean (std. error)
A	1.561	4.48 (1.68)	3.70 (1.07)
B	2.984	2.02 (0.82)	2.52 (0.73)
C	1.582	2.53 (1.27)	2.91 (0.92)
D	4.767	1.89 (0.63)	2.32 (0.61)
E	21.140	4.87 (0.48)	4.66 (0.46)
F	4.659	4.08 (0.93)	3.76 (0.76)
G	2.010	3.10 (1.11)	3.12 (0.83)
H	7.998	1.62 (0.45)	2.00 (0.48)
System	46.701	—	3.20 (0.64)

Figure 1 shows the monthly frequencies of CLI for each hospital, together with the estimated hospital-specific median frequencies and 95% credible intervals. Hospital C had nine missing data points for which the estimated values are shown. All the hospitals except hospital E showed low frequencies of infection with all observed frequencies within the 95% credible intervals. Frequencies of infection for hospital E were higher than were the rest (4.87 infections per 1,000 patient-line days) and it had high monthly 1,000 patient-line days (an average of 1.762 per month). Nonetheless, only one data point fell above the upper limit.

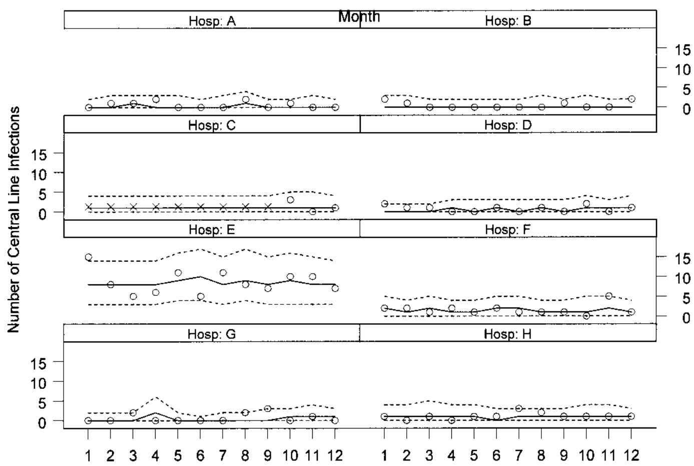

Figure 1. Observed frequencies (circles) of central line infections with hospital median frequency (solid lines) and upper and lower 95% credible limits (dashes).

Method III: System Benchmark

The method for estimating the system benchmark for CLI was calculated from a general statistical model for outcomes (Daniels & Gatsonis, 1999; Normand, Glickman, & Gatsonis, 1997). With this approach, researchers assume that, in addition to the CLI for each hospital occurring at a rate according to a hospital-specific distribution, the hospital CLI rates themselves are determined from a second distribution of rates determined for the system. The analysis of the system benchmark was based on the same data set as was the analysis of hospital-specific benchmarks, but it included two sampling frames: hospitals and system. The first sampling frame was CLI within hospital in which 1,000 patient-line-days was the unit of analysis. This sampling frame is implicit in the analysis of hospital-specific rates. The second sampling frame was the hospital CLI rate within a system in which a hospital was the unit of analysis. The sample for the second frame was thus the eight hospitals.

For the second sampling frame, the distribution of rates was assumed to follow a gamma distribution (a standard distribution for rates). The system rate of CLI and the system-based hospital rates were estimated by standard Bayesian methods for the Poisson-gamma distribution (Bernardo & Smith, 1994), using the same computer programs previously cited.

Results

Table 1 indicates the overall system rate (last row) and the system-based rates for each hospital. The estimated system rate was 3.20 CLI per 1,000 patient-line days. The estimated hospital rates ranged from a low of 2.00 (Hospital H) to a high of 4.66 (Hospital E). In comparing the hospital-specific rates to system-based rates, we found that the latter decreased toward the system rate. This “shrinkage” occurred because the system-based hospital rate was a weighted average of the hospital-specific rate and system rate. This shrinkage reverses the rankings of Hospitals A and F. For the hospital-specific rates, Hospital A had a greater rate than did F (4.48 versus 4.08), but for the system-based rate, the ranking was reversed (3.70 versus 3.76). The small sample size (1.561 1,000 patient-line days) for Hospital A gave less weight to its estimate, allowing the estimate to be pulled toward the system rate. In all but one hospital (Hospital H), the standard error of the system-based estimated rate was less than that of the hospital-specific estimated rate.

Figure 2 shows the monthly observed frequencies of CLI for each hospital (same as in **Figure 1)** together with the estimated system median (50th percentile) frequency and the 95% credible interval. The system-based 95% credible intervals are slightly wider than those for the corresponding hospital-specific intervals, as the system intervals also include both within- and between-hospital variation. The exception is Hospital A: Its small sample size had a hospital-specific credible interval slightly wider (less precise) than the corresponding system-based interval had.

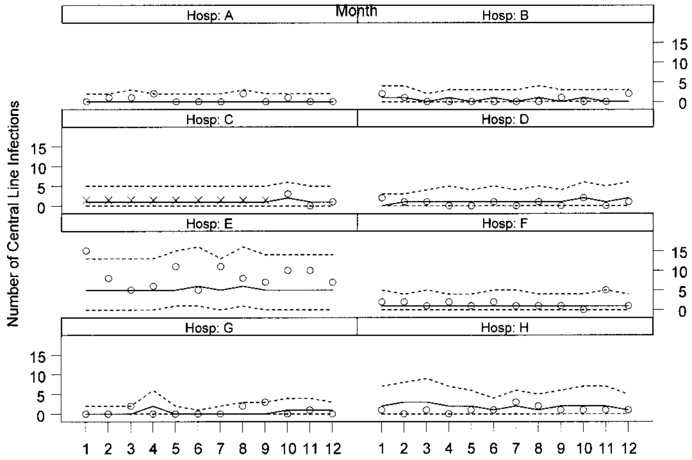

Figure 2. Observed frequencies (circles) of central line infections with system median frequency (solid lines) and upper and lower 95% credible limits (dashes).

All the hospitals showed all observed frequencies within the 95% credible intervals with the exception of one observation for Hospital E. Ten of the 12 observations for Hospital E were in the upper 50th percentile of the intervals. Eight of the 12 observations for Hospital H were in the lower 50th percentile of the intervals. Hospital E had only 1 month outside the benchmarking interval in both hospital-specific and system analyses. However, Hospital E went from 50% in the upper half of the intervals for the hospital-specific analysis to 83% in the upper half for the system analysis.

Discussion

From benchmarks based on the literature, hospital performances were mixed with regard to central line infections. Five of the eight hospitals had hospital-specific rates exceeding 2.2 infections per 1,000 patient-line days, the rate reported by Gastmeier and colleagues (1999) using a large sample of ICU patients. Two hospitals had rates exceeding 3.98, and one had a rate exceeding 4.2, comparable to rates reported by Gowardman and colleagues (1998) in a much smaller sample.

With hospital-specific data, each hospital appeared to be doing reasonably well when compared to its own 95% credible interval. Hospital E showed a single instance of central-line infection above the upper level. Thus, evaluation of each hospital in isolation from the others could result in the conclusion that each hospital was meeting its benchmark. The strength of hospital-specific benchmarking is that hospital type, level of technology, and patient acuity (provided they do not change over the benchmarking period) are controlled.

From the system perspective, using the 95% credible interval indicated a somewhat different circumstance. Hospital E still showed only a single observation above the upper limit, but also showed that 10 of 12 frequencies were in the upper half of the system intervals. That one observation, appearing as single aberration in the hospital-specific analysis, now appeared to be but an instance of a larger pattern of increased frequencies. At the other end of the spectrum, Hospital H had low frequency of infections, with 8 of the 12 observations falling below the system median. Thus, considering eight hospitals as an integrated system, we found all the hospitals performing within the acceptable limits. However, Hospital E had a consistent pattern of higher frequencies of central line infections. The strength of using a multihospital system benchmark is in forcing comparisons among institutions, giving quality indicators a broader examination. The limitation of the multihospital approach is the need to control for possible differences among hospital types, levels of technology, and patient acuities within the system.

No attempt was made to account for patient acuity, technological capacity of the hospital, or variations in risk of the patient populations. These factors should be examined as investigators become more sophisticated in analyses of outcomes. However, the variation in acuity among hospitals should not be large, because only critically ill patients have central lines. Attention to risk adjustment of patients and technological capacity of the hospital are next steps.

Conclusions

Within the U.S. health care system, consumers will continue to demand that health care systems provide data indicating high quality patient care (Millenson, 1997). Benchmarking is a common approach to establishing quality. However, the conclusions drawn from benchmarking depend heavily on whether the benchmark is obtained from the literature, from hospital-specific sources, or from an integrated hospital system. Benchmarking using the literature may appear the simplest, but often a literature-based benchmark is not available, is not sufficiently relevant, or differs in definitions, populations, or clinical practice. Reports in the literature to date do not indicate controls for patient acuity, but Gowardman and colleagues' (1998) study showed a higher CLI rate (4.2 per 1,000 catheter-days) for ICU patients. In our study Hospital E's mean rate exceeded the reported ICU rate. An important but rarely addressed issue in literature-based benchmarking is assessing uncertainty, such as the standard error, in the benchmark itself.

Internal benchmarking is available to hospitals with the relevant databases and statistical expertise, but it can provide an invalid assessment of performance when compared to other institutions. This approach is akin to assessing your chances of winning a car race by looking only at your own speedometer and not noticing the speeds of competing cars.

System benchmarking appears to avoid the pitfalls of these other two methods, but it requires coordinated database resources and sophisticated statistical analyses. System-based benchmarks without adequate adjustments for acuity put hospitals with higher acuity at a disadvantage. Hospitals with smaller censuses may have larger differences between hospital-specific and system-based estimates than do those with larger censuses. These issues are difficult and complex, yet the providing of health care is complex, and the analyses of outcomes reflect that complexity.

References

Aiken, L.H., Sloane, D.M., Lake, E.T., Sochalski, J., & Weber, A.L. (1999). Organization and outcomes of inpatient AIDS care. **Medical Care, 37**, 760-772.

Baggs, J.G., Ryan, S.A., Phelps, C.E., Richeson, J.F., & Johnson, J.E. (1992). The association between interdisciplinary collaboration and patient outcomes in a medical intensive care unit. **Heart and Lung, 21**(1), 18-24.

Bernardo, J.M., & Smith, A.F.M. (1994). **Bayesian theory.** Chichester, U.K.: Wiley.

Byrne, J., Schreiner, S., Rizk, E., & Sokolowski, L. (1998). Clinical quality benchmarking: Here's how to do it — and why. **Trustee, 51**(5, Suppl.),1-4.

Daniels, M.J., & Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. **Journal of the American Statistical Association, 94**(445), 29-42.

Epstein, A.M. (1998). Rolling down the runway. The challenges ahead for quality report cards. **Journal of the American Medical Association, 279**(21), 1691-1696.

Fink, A., Yano, E.M., & Brook, R.H. (1989). The condition of the literature on differences on hospital mortality. **Medical Care, 27**(4), 315-336.

Gastmeier, P., Weist, K., & Ruden, K. (1999). Catheter-associated primary bloodstream infections: Epidemiology and preventive methods. **Infection, 27**(Suppl. 1), 51-56.

Goetz, A.M., Wagener, M.M., Miller, J.M., & Muder, R.R. (1998). Risk of infection due to central versus catheters: Effect of site placement and catheter type. **Infection Control and Hospital Epidemiology, 19**(11), 842-845.

Gowardman, J.R., Montgomery, C., Thirlwell, S., Shewan, J., Idema, A., Larsen, P.D., & Havill, J.H. (1998). Central versus catheter-related bloodstream infections: An analysis of incidence and risk factors in a cohort of 400 patients. **Intensive Care Medicine, 24**(10), 1034-1039.

Knaus, W.A., Draper, E.A., Wagner, D.P., & Zimmerman, J.E. (1986). An evaluation of outcome from intensive care in major medical centers. **Annals of Internal Medicine, 104**(3), 410-418.

Lichtig, L.K., Knauf, R.A., & Milholland, D.K. (1999). Some impacts of nursing on acute care hospital outcomes. **Journal of Nursing Administration, 29**(2), 25-33.

Little, R.J.A., & Rubin, D.B. (1987). **Statistical analysis with missing data.** New York: Wiley.

Millenson, M.L. (1997). **Demanding medical excellence: Doctors and accountability in the information age.** Chicago: University of Chicago.

Mitchell, P.H., Armstrong, S., Simpson, T.F., & Lentz, M. (1989). American association of critical care nurses demonstration project: Profile of excellence in critical care nursing. **Heart and Lung, 18**(3), 219-237.

National Nosocomial Infections Surveillance (NNIS). (1999, June). National Nosocomial Infections Surveillance (NNIS) System Report, Summary from January 1990-May 1999. **American Journal of Infection Control, 27**, 520-532.

Normand, S.T., Glickman, M.E., & Gatsonis, C. (1997). Statistical methods for profiling providers of medical care: Issues and applications. **Journal of the American Statistical Association, 92**(439), 803-814.

Office of Technology and Assessment. (1995). **Hospital financing in seven countries.** Washington, DC: US Government Printing Office.

Paneth, N., Kiely, J.L., Wallenstein, S., Marcus, M., Pakter, J., & Susser, M. (1982). Newborn intensive care and neonatal mortality in low-birth-weight infants. **New England Journal of Medicine, 307**(3), 149-155.

S-PLUS 2000 User's Guide. (1999). Seattle, WA: MathSoft Data Analysis Products Division.

Scott, W.R., Flood, A.B., & Ewy, W. (1979). Organizational determinants of services, quality and costs of care in hospitals. **Milbank Memorial Fund Quarterly, 57**(2), 234-264.

Shortell, S.M., Zimmerman, J.E., Rousseau, D.M., Gillies, R.R., Wagner, D.P., Draper, E.A., Knaus, W.A., & Duffy, J. (1994). The performance of intensive care units: Does good management make a difference? **Medical Care, 32**(5), 508-525.

Spiegelhalter, D., Thomas, A., & Best, N. (1999). **WinBUGS version 1.2 user manual.** Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Whitman, G.R., Davidson, L.J., Rudy, E.B., & Wolf, G.A. (2001). Developing a multi-institutional nursing report card. **Journal of Nursing Administration, 31**(2), 1-7.