Surgical Mortality as an Indicator of Hospital Quality

The Problem With Small Sample Size

Justin B. Dimick, MD
H. Gilbert Welch, MD, MPH
John D. Birkmeyer, MD

Context Surgical mortality rates are increasingly used to measure hospital quality. It is not clear, however, how many hospitals have sufficient caseloads to reliably identify quality problems.

Objective To determine whether the 7 operations for which mortality has been advocated as a quality indicator by the Agency for Healthcare Research and Quality (coronary artery bypass graft [CABG] surgery, repair of abdominal aortic aneurysm, pancreatic resection, esophageal resection, pediatric heart surgery, craniotomy, hip replacement) are performed frequently enough to reliably identify hospitals with increased mortality rates.

Design and Setting The US national average mortality rates and hospital caseloads of the 7 operations were determined using the 2000 Nationwide Inpatient Sample (NIS), and sample size calculations were performed to determine the minimum caseload necessary to reliably detect increased mortality rates in poorly performing hospitals. A 3-year hospital caseload was used for the baseline analysis, and poor performance was defined as a mortality rate double the national average.

Main Outcome Measure Proportion of hospitals in the United States that performed more than the minimum caseload for each operation.

Results The national average mortality rates for the 7 procedures examined ranged from 0.3% for hip replacement to 10.7% for craniotomy. Minimum hospital caseloads necessary to detect a doubling of the mortality rate were 64 cases for craniotomy, 77 for esophageal resection, 86 for pancreatic resection, 138 for pediatric heart surgery, 195 for repair of abdominal aortic aneurysm, 219 for CABG surgery, and 2668 for hip replacement. For only 1 operation did the majority of hospitals exceed the minimum caseload, with 90% of hospitals performing CABG surgery having a caseload of 219 or higher. For the remaining operations, only a small proportion of hospitals met the minimum caseload: craniotomy (33%), pediatric heart surgery (25%), repair of abdominal aortic aneurysm (8%), pancreatic resection (2%), esophageal resection (1%), and hip replacement (<1%).

Conclusion Except for CABG surgery, the operations for which surgical mortality has been advocated as a quality indicator are not performed frequently enough to judge hospital quality.

JAMA. 2004;292:847-851 www.jama.com

PATIENTS AND POLICY MAKERS INcreasingly use rates of surgical mortality to assess hospital performance. New York and Pennsylvania have long-standing systems for tracking and publicly reporting risk-adjusted mortality rates after cardiac surgery[1,2]; California and New Jersey have more recently adopted this approach.[3,4] The Leapfrog Group, a large coalition of employers and purchasers, has made surgical mortality rates one of the criteria for "evidence-based referral" for cardiac procedures.[5] As part of its broader efforts to develop a core set of quality indicators, the Agency for Healthcare Research and Quality (AHRQ) has recently endorsed the use of surgical mortality rates for 7 surgical procedures including repair of abdominal aortic aneurysm, esophageal resection, and hip replacement.[6]

However, there are 2 reasons to question whether rates of surgical mortality can reliably detect quality problems. First, the targeted operations are infrequently performed at individual hospitals. Second, the mortality rates for many of these procedures are often relatively low. Small samples and low event rates combine to limit the statistical power of a comparison between an individual hospital and a population-based bench-

Author Affiliations: VA Outcomes Group, Department of Veterans Affairs Medical Center, White River Junction, Vt (Drs Dimick and Welch); Center for the Evaluative Clinical Sciences, Dartmouth Medical School, Hanover, NH (Drs Dimick and Welch); and Michigan Surgical Collaborative for Outcomes Research and Evaluation (M-SCORE), Department of Surgery, University of Michigan Medical Center, Ann Arbor (Drs Dimick and Birkmeyer).
Corresponding Author: Justin B. Dimick, MD, VA Outcomes Group 111B, Department of Veterans Affairs Medical Center, 215 N Main St, White River Junction, VT 05009 (justin.b.dimick@dartmouth.edu).

Table 1. Characteristics of the 7 Included Operations in the United States, 2000

Characteristic	Repair of Abdominal Aortic Aneurysm	CABG Surgery	Craniotomy	Esophageal Resection	Hip Replacement	Pancreatic Resection	Pediatric Heart Surgery
No. of hospitals performing operation	2485	1036	1600	1717	3445	1302	458
National average mortality rate, %	3.9	3.5	10.7	9.1	0.3	8.3	5.4
Annual hospital caseloads, median (IQR)	30 (17-55)	491 (274-852)	12 (4-30)	5 (2-10)	24 (9-58)	8 (4-24)	4 (1-50)

Abbreviation: CABG, coronary artery bypass graft; IQR, interquartile range.

mark. The practical implication of limited power is that patients and policymakers may not identify a hospital with quality problems. Although the general problem of failing to detect important differences—type II error—is well recognized in the context of clinical trials, it is often overlooked in quality measurement.[7]

This study was designed to explore this problem for the 7 surgical procedures suggested for mortality measurement by the AHRQ. Using data from the Nationwide Inpatient Sample (NIS), we determined the national average mortality rate for each procedure and the number of cases performed in each hospital. We then estimated the minimum sample size needed to identify a poorly performing hospital as significantly different from the national average mortality rate. Finally, we determined the proportion of US hospitals that exceed this minimum caseload—ie, those hospitals for which mortality would reliably reflect quality.

METHODS

Data Source

The data are from the 2000 NIS maintained by the AHRQ as part of the Healthcare Cost and Utilization Project.[8] The NIS is a database of all discharges from a nationally representative sample of 994 hospitals (randomly selected within strata for region, number of hospital beds, teaching status, urban vs rural location, and hospital ownership) containing data for approximately 20% of all acute care hospitalizations in the United States. Hospital weights were used to generate estimates of mortality rates and caseload distributions that represent all hospitals in the United States. Because the NIS includes different hospitals each year, we were not able to directly determine hospital caseloads over several years. We therefore assumed that hospital caseloads are constant over time and estimated 3- and 5-year caseloads using the 2000 NIS data.

Selection of Operations

We examined the 7 surgical procedures for which mortality has been advocated as a performance measure by the AHRQ Inpatient Quality Indicators[6]: coronary artery bypass graft (CABG) surgery, repair of abdominal aortic aneurysm, pancreatic resection, esophageal resection, pediatric heart surgery, craniotomy, and hip replacement. For 6 of the 7 operations, discharges were identified in the NIS database using the appropriate combination of *International Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*)[9] procedure and diagnostic codes suggested by the AHRQ. For craniotomy, the AHRQ selection criteria were based on a diagnosis related group code 01, which includes less-extensive procedures (eg, sinus surgery, shunt placement). We therefore restricted the analysis to discharges with an *ICD-9-CM* procedure code for craniotomy.

Analytic Approach

The analysis had 4 steps. First, we used the NIS data to determine our benchmark, ie, the national average mortality rate for each procedure. Mortality was defined as any death during the index hospital stay.

Second, we performed a sample size calculation for each procedure to determine the minimum caseload necessary to reliably detect increased mortality in a poorly performing hospital. For the baseline analysis, we defined poorly performing hospitals as having a mortality rate of twice our benchmark (ie, the effect size for the sample size calculation was the difference between the national average mortality rate and twice the national average mortality rate).

Sample size calculations were based on 1-sample, 1-sided tests ($\alpha = .05$) with a power of 80%. Although unusual, 1-sample, 1-sided tests are appropriate for the task: 1-sample tests because we are interested in detecting whether an individual hospital is significantly different than a population benchmark, and 1-sided tests because we are only interested in determining whether the hospital mortality is higher than the benchmark. The effect of both assumptions is to reduce the minimum caseload necessary for each procedure; thus, the assumptions are conservative given our question. Sample size calculations were performed using STATA version 8.0 (STATA Corp, College Station, Tex).

Third, we determined the proportion of hospitals that met or exceeded the minimum caseload in the NIS. The numerator of each proportion was the number of hospitals that met or exceeded the minimum caseload over a 3-year period. The denominator included all hospitals performing at least 1 procedure.

Finally, because hospital performance may be measured over different time periods, we conducted sensitivity analyses varying the period of observation from a low of 1 year to a high of 5 years. We also varied the definition of poor performance and repeated the analysis using a more subtle increase of 1.5 times the national average mortality rate (our benchmark).

RESULTS

TABLE 1 shows the national average mortality rates for the 7 procedures exam-

ined, which ranged from 0.3% for hip replacement to 10.7% for craniotomy. Table 1 also shows that annual caseloads of individual hospitals varied widely. The median caseloads ranged from 4 cases per hospital for pediatric heart surgery to 491 for CABG surgery. The variability is also evident within operations. For example, the caseloads for repair of abdominal aortic aneurysm ranged from 1 to 199 per hospital.

The minimum hospital caseload necessary to detect a mortality rate of twice the national benchmark was inversely related to the operative mortality rate of each procedure. The minimum caseloads varied from 64 cases for craniotomy to 2668 cases for hip replacement. The minimum caseloads for other operations were as follows: esophageal resection (77), pancreatic resection (86), pediatric heart surgery (138), repair of abdominal aortic aneurysm (195), and CABG surgery (219).

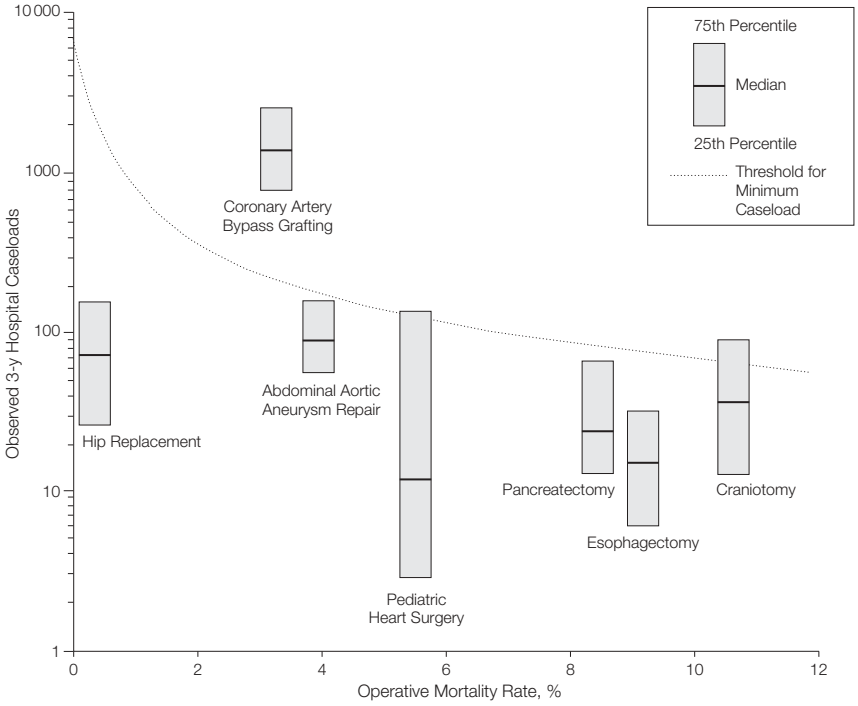
For only 1 operation did the majority of hospitals exceed the minimum caseload: 90% of hospitals performing CABG surgery had a caseload of 219 or higher. For the remaining operations, only a small proportion of hospitals met the minimum caseload: craniotomy (33%), pediatric heart surgery (25%), repair of abdominal aortic aneurysm (8%), pancreatectomy (2%), esophagectomy (1%), and hip replacement (<1%). The FIGURE presents a detailed view of the data, showing the distribution of actual hospital caseloads relative to the minimum caseload needed to detect a doubling of the operative mortality rate.

The impact of changing the number of years of data and altering the definition of poor performance is shown in TABLE 2. Even after increasing the sample size by using 5-year hospital caseloads, CABG surgery was still the only operation with more than half of US hospitals having more than the minimum caseload. Changing the definition of poor performance to 1.5 times the benchmark (ie, a 50% increase above the national average mortality rate) dramatically decreased the number of hospitals having more than the minimum caseload. Under these conditions, only 54% of US hospitals performing CABG surgery met the minimum caseload in a 3-year period.

COMMENT

Although the problem of small sample size has received considerable attention in the context of clinical trials, it has re-

Figure. Distribution of Actual Hospital Caseloads and the Minimum Caseload Needed to Detect a Doubling of the Mortality Rate



Area above the dotted line indicates the minimum caseloads exceeded for a given mortality rate.

Table 2. Sensitivity Analysis: Proportion of US Hospitals Meeting or Exceeding the Minimum Caseloads After Varying the Number of Years and the Definition of Poor Performance

Surgical Procedure	Poor Performance: Mortality Rate Increase Over Benchmark	Minimum Caseload	Hospitals Meeting or Exceeding Minimum Caseload, %		
			1 Year of Data	3 Years of Data	5 Years of Data
CABG surgery	Double	219	61	90	94
	1.5 times	744	12	54	78
Craniotomy	Double	64	10	33	48
	1.5 times	230	<1	6	15
Pediatric heart surgery	Double	138	9	25	31
	1.5 times	497	1	7	17
Repair of abdominal aneurysm	Double	195	<1	8	18
	1.5 times	668	None	None	1
Pancreatic resection	Double	86	None	2	3
	1.5 times	304	None	None	None
Esophageal resection	Double	77	None	1	4
	1.5 times	275	None	None	None
Hip replacement	Double	2668	None	<1	<1
	1.5 times	9512	None	None	None

Abbreviation: CABG, coronary artery bypass graft.

ceived less consideration in quality measurement. The ability to detect a difference between the mortality rate at an individual hospital and a benchmark rate is dependent on both the baseline mortality rate and the number of cases performed. For mortality to be a useful measure of quality, the procedure must both have a relatively high mortality rate and be performed frequently. Our findings suggest that CABG surgery fulfills these criteria. For the other 6 operations in our analysis, however, less than half of hospitals in the United States perform enough cases to detect a doubling of the mortality rate. For these procedures, other approaches to measuring quality will be required.

The problem of continuing to use rates of surgical mortality as an indicator of hospital quality is perhaps most pronounced for hospitals with truly poor performance. These hospitals are falsely reassured that their performance is "average" and therefore have less incentive to improve. Payers are falsely reassured that they are buying a good product and miss an opportunity to steer patients away from poorly performing hospitals through selective contracting or other mechanisms. Patients are falsely reassured that they are choosing a safe hospital.

Regarding the generalizability of our findings, the sample used in our analysis included only 20% of hospitals in the United States. However, the hospitals were chosen as a stratified random sample that is specifically intended to represent all US hospitals.[8] Furthermore, our analysis was limited to 7 surgical procedures, which were selected because of their inclusion in the AHRQ Inpatient Quality Indicators.[6] These operations represent a wide range of mortality rates and hospital caseloads. Although other operations could have been assessed, it is difficult to identify any that are both common enough and sufficiently high-risk for mortality to be a useful quality measure.

Regarding the assumptions used in our calculations, in our baseline analysis we sought to detect a doubling of the mortality rate. Few would argue that this increase is not clinically significant. In fact, both patients and physicians would likely be interested in detecting more subtle differences in performance. Based on our sensitivity analysis, it is clear that the usefulness of mortality rates markedly declines when attempting to detect a mortality rate of 1.5 times the benchmark. Furthermore, we used 3-year hospital caseloads in our baseline analysis. Increasing the period of observation to 5 years in our sensitivity analysis had little effect on our findings. Mortality rates based on longer periods of observation are less relevant to current performance, since surgical staff and practices may change over time.

We admittedly did not explicitly consider risk adjustment in our analysis. Indeed, over the past 2 decades, the question of whether mortality rates are useful measures of quality has focused largely on issues of risk adjustment. However, while we acknowledge that adjusting for differences between hospitals is of crucial importance, we believe that concerns of adequate sample size must be addressed first. Without sufficient sample size, even perfect risk adjustment does not matter.

Given the limited usefulness of procedure-specific mortality rates, it is worth considering additional approaches to judging surgical quality. The first alternative approach would be to increase the number of observations by combining operations to produce an aggregate mortality rate. Perhaps the most visible example of this approach is the National Surgical Quality Improvement Program used in Veterans Affairs hospitals.[10] This approach, however, lacks information on the quality of care for individual procedures. For instance, a hospital's high mortality with carotid endarterectomy may be masked by the very low mortality of more common operations such as laparoscopic cholecystectomies, hernia repairs, and appendectomies. When determined from heterogeneous groups of procedures, an increased mortality rate does not provide clear guidance about where to focus efforts at quality improvement.

A second approach would be to focus on other outcomes. To be useful, these other outcomes must occur more frequently than mortality. Postoperative complications often meet this requirement, but the clinical severity of complications is extremely broad (eg, from superficial wound infection to ventilator-associated pneumonia), and there is currently no standard approach to their classification for quality measurement. Measures that are based on patient-oriented outcomes such as quality of life, time until the return to work, or patient satisfaction offer several advantages. Unlike mortality, all patients undergoing a given operation experience these outcomes—making it feasible to detect clinically important differences with smaller sample sizes. The usefulness of these measures, however, depends on the availability of detailed clinical data. Such data are not widely available and are often not practical on a large scale given the expense of data collection.

The third approach would be to focus on indirect measures of quality, including processes of care and procedural volume. Focusing on processes of care has been standard for quality measurement for medical conditions (eg, β-blockers for patients after myocardial infarction). Although this approach could be applied to surgery in principle, the number of specific processes of care that would be useful for this purpose is limited. One quality indicator that has received significant attention recently is hospital or surgeon volume; both have been shown to be associated with lower mortality rates for many complex operations.[11] Information on hospital volume is easy and inexpensive to obtain. However, volume is often a poor predictor of performance for individual hospitals or surgeons.

No single quality measure will be appropriate for all operations. Mortality may work well for CABG surgery but will be too imprecise for use with other procedures, such as pancreatic resection. Policy makers should consider sample size in selecting the best qual-

ity measure for specific procedures, particularly when data are used for public reporting. Otherwise, they run the risk of mislabeling hospitals and misinforming patients.

Author Contributions: Dr Dimick had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analyses.
Study concept and design: Dimick, Welch, Birkmeyer.
Acquisition of data: Dimick.
Analysis and interpretation of data: Dimick, Welch, Birkmeyer.
Drafting of the manuscript: Dimick, Welch.
Critical revision of the manuscript for important intellectual content: Dimick, Welch, Birkmeyer.
Statistical analysis: Dimick, Welch.
Obtained funding: Welch, Birkmeyer.
Administrative, technical, or material support: Dimick, Birkmeyer.
Study supervision: Welch, Birkmeyer.
Funding/Support: Dr Dimick was supported by a Veterans Affairs Special Fellowship Program in Outcomes Research. This study was also supported by a Research Enhancement Award from the Department of Veterans Affairs to investigate the harms from excessive medical care.
Role of the Sponsor: The Department of Veterans Affairs provided salary support to the investigators but had no role in the design and conduct of the study; the collection, analysis, and interpretation of the data; the preparation of the data; or the preparation, review, or approval of the manuscript.
Disclaimer: The views expressed herein do not necessarily represent the views of the Department of Veterans Affairs or the federal government.

REFERENCES

1. *Coronary Artery Bypass Surgery in New York State: 1989-1991*. Albany: New York State Dept of Health; 1992.
2. *A Consumer Guide to Coronary Artery Bypass Graft Surgery*. Harrisburg: Pennsylvania Health Care Cost Containment Council; 1991.
3. *Coronary Artery Bypass Graft Surgery in New Jersey: 1994-1995*. Trenton: New Jersey Dept of Health and Senior Services; 1997.
4. *The California Report on Coronary Artery Bypass Graft Surgery: 1997-1998 Hospital Data*. Sacramento: Pacific Business Group on Health and the California Office of Statewide Health Planning and Development; 2001.
5. Milstein A, Galvin RS, Delbanco SF, Salber P, Buck CR Jr. Improving the safety of health care: the Leapfrog initiative. *Eff Clin Pract*. 2000;3:313-316.
6. *AHRQ Quality Indicators—Guide to Inpatient Quality Indicators: Quality of Care in Hospitals—Volume, Mortality, and Utilization*. Rockville, Md: Agency for Healthcare Research and Quality; 2002. AHRQ Publication 02-R0204.
7. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. *N Engl J Med*. 1978;299:690-694.
8. Healthcare Cost and Utilization Project (HCUP-9). *Nationwide Inpatient Sample, Release 9*. Rockville, Md: Agency for Healthcare Research and Quality; 2000.
9. *International Classification of Diseases, Ninth Revision, Clinical Modification*. Washington, DC: Public Health Service, US Dept of Health and Human Services; 1988.
10. Khuri SF, Daley J, Henderson WG. The comparative assessment and improvement of the quality of surgical care in the Department of Veterans Affairs. *Arch Surg*. 2002;137:20-27.
11. Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon volume and operative mortality in the United States. *N Engl J Med*. 2003;349:2117-2127.

The opposite of a correct statement is a false statement. But the opposite of a profound truth may well be another profound truth.
—Niels Bohr (1885-1962)