Does Surgical Quality Improve in the American College of Surgeons National Surgical Quality Improvement Program

An Evaluation of All Participating Hospitals

Bruce L. Hall, MD, PhD, MBA, FACS,†‡§ Barton H. Hamilton, PhD,§ Karen Richards, BS,¶ Karl Y. Bilimoria, MD, MS,‖ Mark E. Cohen, PhD,¶ and Clifford Y. Ko, MD, MS, MSHS, FACS**¶*

Background/Objective: The National Surgical Quality Improvement Program (NSQIP) has demonstrated quality improvement in the VA and pilot study of 14 academic institutions. The objective was to show that American College of Surgeons (ACS)-NSQIP helps all enrolled hospitals.

Methods: ACS-NSQIP data was used to evaluate improvement in hospitals longitudinally over 3 years (2005–2007). Improvement was defined as reduction in risk-adjusted "Observed/Expected" (O/E) ratios between periods with risk adjustment held constant. Multivariable logistic regression-based adjustment was performed and included indicators for procedure groups. Additionally, morbidity counts were modeled using a negative binomial model, to estimate the number of avoided complications.

Results: Multiple perspectives reflected improvement over time. In the analysis of 118 hospitals (2006–2007), 66% of hospitals improved risk-adjusted mortality (mean O/E improvement: 0.174; $P < 0.05$) and 82% improved risk adjusted complication rates (mean improvement: 0.114; $P < 0.05$). Correlations between starting O/E and improvement (0.834 for mortality, 0.652 for morbidity), as well as relative risk, revealed that initially worse-performing hospitals had more likelihood of improvement. Nonetheless, well-performing hospitals also improved. Modeling morbidity counts, 183 hospitals (2007), avoided ~9598 potential complications: ~52/hospital. Due to sampling this may represent only 1 of 5 to 1of 10 of the true total. Improvement reflected aggregate performance across all types of hospitals (academic/community, urban/rural). Changes in patient risk over time had important contributions to the effect.

Conclusions: ACS-NSQIP indicates that surgical outcomes improve across all participating hospitals in the private sector. Improvement is reflected for both poor- and well-performing facilities. NSQIP hospitals appear to be avoiding substantial numbers of complications- improving care, and reducing costs. Changes in risk over time merit further study.

(*Ann Surg* 2009;250: 363–376)

From the *Department of Surgery, John Cochran Veterans Affairs Medical Center, St Louis, MO; †Washington University Center for Health Policy, St Louis, MO; ‡Department of Surgery, Washington University in Saint Louis School of Medicine, St Louis, MO; §Olin Business School at Washington University in St Louis, St Louis, MO; ¶Division of Research and Optimal Patient Care, American College of Surgeons, Chicago, IL; ‖Department of Surgery, Northwestern University School of Medicine, Chicago, IL; and **Department of Surgery, University of California Los Angeles School of Medicine, Los Angeles, CA.

Supported by the Center for Health Policy, Washington University in Saint Louis, director William Peck, MD (to B.L.H.) and also by the American College of Surgeons Clinical Scholars in Residence program (to K.Y.B.).

The ACS NSQIP and the hospitals participating in the ACS NSQIP are the source of the data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors. This study does not represent the views or plans of the ACS or the ACS NSQIP.

Reprints: Bruce L. Hall, MD, PhD, MBA, Campus Box 8109, 660 South Euclid Ave, St. Louis, MO 63110. E-mail: hallb@ wustl.edu.

ISSN: 0003-4932/09/25003-0363
DOI: 10.1097/SLA.0b013e3181b4148f

The National Surgical Quality Improvement Program (NSQIP) was developed in the 1990s in the Veterans Health Administration and led to marked improvement in surgical quality. Mortality and morbidity rates declined, patient satisfaction improved, and lengths of stay decreased.[1,2] In 2001 to 2004, with funding from the Agency for Healthcare Research and Quality, a pilot study outside the VA, the Patient Safety in Surgery Study, was performed which demonstrated that NSQIP was feasible to implement in the private sector, and resulted in aggregate reduction of postoperative morbidity.[3] The American College of Surgeons NSQIP (ACS-NSQIP) was subsequently opened to the private sector by subscription after 2004. The ACS-NSQIP collects data and reports risk adjusted surgical outcomes. It is the only multispecialty, clinically based, prospectively collected, quality improvement (QI) program for the profession of surgery, and its utility has been shown over years of implementation. The program has grown in the private sector since inception, and continues to grow. It now includes >200 hospitals varying in size, location, and teaching status. The objective of this study was to show whether the ACS-NSQIP helps enrolled hospitals improve surgical quality over time.

METHODS

The NSQIP general approach to data collection and performance evaluation has been described previously.[1–8] In brief, the program has traditionally focused on general and vascular surgery (outside of the VA) although a multispecialty approach is now available. The program's strengths include reliance on clinical data (not administrative) abstracted from the medical record by a trained data expert. The program focuses on 30-day outcomes (whether or not a patient has been discharged from their initial admission) via direct ascertainment of the 30-day time point. Outcomes include 21 rigorously defined morbidities (including the following categories: wound, respiratory, urinary tract, central nervous system, cardiac, and 5 others), as well as mortality. Eligible cases include major general and vascular cases under general/spinal/epidural anesthesia, subject to eligibility and accrual limits. Cases are sampled in a systematic, temporal fashion. A critical feature of the program has been that data collection is coordinated by a dedicated full time nurse or trained health information expert, who is specifically trained in NSQIP methods and data field definitions, who is regularly audited, and who maintains a degree of separation from individual surgeons. Specific materials describing the qualifications, training, and auditing of these personnel, as well as data definitions and data collection protocols, are available online from the ACS NSQIP website.[9] A prominent aspect of the approach is regular assessment of interrater reliability. As a result of multiple reinforcing approaches, data integrity within the program has been excellent and consistently improving as well. For instance, interrater reliability audits revealed that in 2005 total disagreements across the program were at 3.15% (for nearly 40,000 audited fields), and by 2008 total disagreements were at 1.60% (>140,000 audited fields).

In 2008, only 2 data fields (>135) had interrater disagreement >5% (personal communication, M. Shiloach at ACS; all work in preparation for publication).

As the ACS-NSQIP has grown over time, the number of participating hospitals has increased. At the end of 2004, 18 hospitals participated. At the end of 2005, there were 37 hospitals, for 2006, 121; and for 2007, 183. Currently, in 2009, there are more than 200 participating institutions. We analyzed the performance of hospitals in the program in 2005, 2006, and 2007, specifically examining changes from 2005 to 2006, 2006 to 2007, and 2005 to 2007. The requirement for participation across periods limited the number of hospitals analyzed, as reported below. We analyzed blinded, preexisting data with all patient identifiers removed, and with institutions identified only by a random code.

Risk adjustment was performed using multivariable logistic regression against outcomes of "any morbidity" or mortality, using the available independent factors collected by NSQIP.[9] We included in our adjustment an indicator for surgical procedure groups, which is not a feature of the standard NSQIP approach.[10] Estimated risk per patient was generated from the resulting coefficients applied to each patient's characteristics. The "Expected" number of events for an institution was calculated by summing all patients' risk estimates for the institution and time period. The actual "Observed" number of events was used to create an "Observed/Expected" (O/E) ratio, and confidence intervals (CI) were generated. A ratio >1.0 indicates performance worse than expected, while an O/E <1.0 indicates performance better than expected. The NSQIP has in the past used 90% CIs on mortality and 99% on morbidity. For this work, in which models differ somewhat from the classic NSQIP approach (for instance by inclusion of category indicators, etc.) we calculated 95% CIs for both, which is a statistical standard. If an institution's O/E is >1 and CI excludes 1; they are a "bad" outlier. If the O/E is <1 and CI excludes 1, they are a "good outlier." O/E ratios were generated for every institution and year, and for any morbidity and mortality. "Change in O/E" between periods was derived by subtracting the initial O/E from the later O/E: a change less than zero indicates improvement. Changes in O/E ratios are presented as absolute reductions in O/E ("O/E units"). For volume-weighted calculations, results for each institution were weighted by their proportion of cases accrued nationally for the year.

Constant Risk Adjustment

We took several approaches to examine change in performance over time. Normally, the NSQIP recalculates risk adjustment every reporting period. This has programmatic advantages, but creates challenges for longer term assessment. In contrast, in our first approach, we evaluated hospitals over more than one period holding risk adjustment constant. The risk adjustment model was based on data for calendar year 2005. We then applied the same risk adjustment algorithm to program data for 2005, 2006, and 2007. This analysis is based on "opportunities for improvement": each hospital participating for 2 or more periods can demonstrate a change in performance. We report a variety of statistics and measures based on this approach, including considerations of outlier status.

Constant Patient Population

In a second approach to evaluating change over time, we held the patient population constant instead of the risk adjustment. In this approach, a "best model" of risk adjustment was created for each year using patient data for each year but only from the set of hospitals originally present in 2005, as per the methods described above. This helps eliminate confounding by "new" potentially dissimilar institutions entering the NSQIP. Then, the patient population from 2005 was passed through each model. This enabled commentary on whether this set of hospitals appeared to be improving their care of the identical patient population over time. This also allowed decomposition of observed changes into patient mix and model effects.

Complications Avoided

In 2 final approaches, we examined outcome events avoided, using 2 structures. First, we focused on the traditional NSQIP outcomes of any morbidity and "mortality." For this structure, we used 2005 risk adjustment applied to each subsequent year, but subtracted observed from expected figures to estimate the events avoided. In a second approach, we examined complication counts. As mentioned, the NSQIP has traditionally evaluated and reported an "any-30 days-morbidity" end point. However, since 40% to 50% of patients with complications experience multiple complications; examining any morbidity as an "all or none" measure could conceal significant improvement and large numbers of complications avoided. For instance, a patient with 2 complications would have to avoid both to exit the any morbidity group. If complications for this patient were reduced from 2 to 1, there would be no change in their any morbidity status. Therefore, we also modeled morbidity counts (not any morbidity), by estimating a zero-inflated, negative binomial model while still adjusting for demographics, comorbidities, indicators for procedure codes, and work relative value units. This modeling was based on data for 2005, 2006, and 2007, again holding risk adjustment constant (2005 coefficients). Expected minus observed counts were calculated. Since these analyses did not require subtracting over 2 time periods, all participating institutions could be analyzed in each year, as reported below.

Statistical analyses were performed using Stata 10 (StataCorp, LP, College Station, TX), Microsoft Excel 2003 (Microsoft Corp., Seattle, WA), or StatsDirect 2.6.6 (StatsDirect Ltd, available at: www.statsdirect.com). Meta analysis of relative risk was performed in StatsDirect using a random effects (DerSimonian-Laird) model. Data were analyzed as paired wherever appropriate.

RESULTS

At the end of 2007 there were 183 ACS-NSQIP hospitals, but only 118 had been in the program for both 2006 and 2007. These institutions accrued 112,069 cases in 2006 (mean: 950, SD: 498) and 155,058 in 2007 (mean: 1314, SD: 440). At the end of 2006, there were 35 hospitals that had been in the program for both 2005 and 2006. These institutions accrued 33,124 cases in 2005 (mean: 946, SD: 611), 49,531 in 2006 (mean: 1415, SD: 433), and 48,719 in 2007 (mean: 1392, SD: 518). At the end of 2005 there were only 11 hospitals that had been in the program during 2004. Therefore, analyses were not extended back to 2004–2005. On the basis of these limitations, we report results for 2 institutional groupings: changes for 118 institutions 2006–2007 and changes for 35 institutions 2005–2006, 2006–2007, and 2005–2007. Table 1 presents summary statistics by year for the analyzed cases, including note of unadjusted morbidity rates by year and group of 12.41%, 12.28%, 11.98%; and unadjusted mortality rates of 1.66%, 1.85%, and 1.77%. Coefficients and odds ratios for the 2005 risk adjustment model, which is the main basis of the investigation, are shown in Table 2. Risk adjustment coefficients for 2006 and 2007, referred to in some results, are not shown.

Holding Risk Adjustment Constant

Aggregate results for 10 hospital groupings/time periods are presented in Table 3.

Net Improvement: 118 Institutions Present From 2006 to 2007

A total of 118 institutions were examined for change between periods 2006 and 2007 (Table 3). Regarding morbidity, 97 of 118 (82%) of institutions improved; for mortality 78 of 118 (66%)

TABLE 1. Summary Statistics for Populations by Year

	2005	2006	2007
Institutions	35	118	118
Cases	33,124	112,069	155,058
30 d morbidity (unadjusted)	12.41%	12.28%	11.98%
30 d mortality (unadjusted)	1.66%	1.85%	1.77%
Demographics			
Age (yr)	54 ± 17	55 ± 17	55 ± 17
Male	41.71%	41.91%	42.85%
Race white	69.08%	68.66%	71.94%
NSQIP preoperative risk factors			
Ascites	1.27%	1.64%	1.91%
Chemotherapy	1.41%	1.26%	1.27%
Coma	0.09%	0.11%	0.07%
Current pneumonia	0.67%	0.67%	0.58%
Cerebrovascular accident with deficit	2.54%	2.40%	2.50%
Cerebrovascular accident without deficit	1.85%	1.91%	2.01%
Diabetes (any)	13.41%	13.82%	14.64%
Dialysis	2.05%	2.08%	2.51%
DNR status	1.12%	0.64%	0.61%
Dyspnea (any)	10.87%	12.02%	11.87%
>2 drinks alcohol/wk	2.99%	2.68%	2.54%
Hemiplegia	1.17%	1.03%	0.99%
Angina	0.84%	0.92%	0.86%
Congestive heart failure	1.10%	1.17%	1.02%
Chronic obstructive pulmonary disease	4.23%	4.31%	4.65%
Myocardial infarction	0.77%	0.78%	0.69%
Peripheral vascular disease	4.49%	4.46%	4.60%
Transient ischemic attack	2.80%	2.73%	2.98%
Hypertension (requiring medication)	42.57%	43.73%	45.10%
Impaired sensorium	1.07%	1.04%	0.89%
Sepsis	8.27%	10.36%	9.59%
Previous cardiac surgery	6.17%	6.33%	6.23%
Previous percutaneous coronary intervention	4.57%	5.24%	5.35%
Radiotherapy	1.04%	0.82%	0.86%
Renal failure	0.59%	0.67%	0.60%
Rest pain	2.64%	2.57%	2.56%
Tobacco use	21.69%	20.87%	20.87%
Steroid use	3.91%	3.33%	3.59%
Transfusion	0.48%	0.50%	0.42%
Ventilator dependent	1.14%	1.16%	1.11%
Wound infection	4.56%	4.88%	5.10%
Weight loss (10 in 6 mo)	3.27%	2.76%	2.73%
Operative information			
Wound class clean	47.16%	45.13%	49.46%
ASA class I–II	60.57%	58.27%	55.51%
Emergency case	13.33%	14.24%	13.15%
Vascular case	12.81%	12.26%	13.81%
Work RVUs	14.47	14.35	15.92
CPT Procedure Families*			
cp2	0.18%	0.14%	0.23%
cp3	10.30%	10.17%	10.59%
cp4	2.33%	2.00%	2.22%
cp5	6.78%	5.32%	5.77%
cp6	0.53%	0.48%	0.55%
cp7	2.57%	2.54%	2.70%
cp8	7.79%	8.44%	9.34%
cp9	4.91%	4.47%	4.76%
cp10	16.33%	16.84%	17.61%
cp11	1.58%	1.59%	1.82%
cp12	17.96%	19.69%	13.27%
cp13	0.51%	0.45%	0.44%
cp14	14.70%	14.84%	15.77%
cp15	0.16%	0.11%	0.09%
cp16	0.13%	0.12%	0.13%
cp17	11.48%	11.20%	12.84%
cp18	0.40%	0.32%	0.36%
cp19	0.18%	0.20%	0.22%

*Procedure families as per Hall et al.[10]

improved. The mean absolute change in O/E was improvement of −0.114 for morbidity, −0.174 for mortality, both highly significant. Weighted by institution volume, the mean absolute change in O/E was improvement of −0.113 for morbidity, −0.163 for mortality. As an example, Figure 1 shows the distribution of O/E changes for morbidity in this group and period. The mean change is negative (improvement) and the population is asymmetrically skewed toward improvement.

The correlation between 2006 O/E and change in O/E for the period was −0.652 for morbidity and −0.834 for mortality, reflecting a relationship between starting assessment being bad, and subsequent improvement. However, the correlation between 2006 O/E and 2007 O/E was 0.766 for morbidity and 0.372 for mortality, reflecting some persistence of good/bad status. Relative starting positions matter, but changes take place.

Net Improvement: 35 Institutions Present From 2005 to 2007

Thirty-five institutions were examined over 2 years 2005–2007 (Table 3). For the change in the first year (2005–2006), regarding morbidity, 19 of 35 (54%) of institutions improved and for mortality 12 of 35 (34%) improved. The mean absolute change in O/E for the period was improvement for morbidity (mean change: −0.040) but worsening for mortality (mean change: +0.124). However, neither was significant. Weighted by institution volume, the mean absolute change in O/E for the period was −0.054 for morbidity and 0.1224 for mortality, again not significant. The correlation between 2005 O/E and change in O/E for the period was −0.693 for morbidity and −0.503 for mortality, again reflecting a tendency for bad performers to improve. Still, the correlation between 2005 O/E and 2006 O/E was 0.501 for morbidity and 0.473 for mortality, again reflecting some persistence of assessment.

In the second year examined (2006–2007), regarding morbidity, 31 of 35 (89%) of institutions improved while for mortality 28 of 35 (80%) improved: both dramatically increased rates of improvement compared with the first year. The mean absolute change in O/E for the period was improvement of −0.123 for morbidity and −0.229 for mortality, both highly significant. Weighted by institution volume, the mean absolute change in O/E for the period was improvement of −0.115 for morbidity and −0.198 for mortality, remaining highly significant. The correlation between 2006 O/E and change in O/E for the period was −0.529 for morbidity and −0.708 for mortality. Furthermore, the correlation between 2005 O/E and change in O/E for the period was −0.322 for morbidity and −0.373 for mortality. Thus, even a relatively poor assessment reaching back 2 years appears correlated with improvement. However, the correlation between change in O/E (2006–2007) and change in O/E (2005–2006) was −0.083 for morbidity and −0.334 for mortality—indicating some difficulty with perpetuating improvement.

These 35 institutions can also be assessed over the combined 2-year period (2005–2007). Regarding morbidity, 31 of 35 (89%) of institutions improved; for mortality 20 of 35 (57%) improved. The mean change in O/E for the period was a significant improvement for morbidity (mean change: −0.162, $P < 0.001$), and for mortality a trend toward improvement not reaching significance (mean change: −0.105, $P = 0.14$). Weighted by institution volume, the mean change in O/E for the period was −0.156 for morbidity and −0.066 for mortality, both confirmatory results. Interestingly, the correlation between 2005 O/E and change in O/E in 2005–2007 was −0.791 for morbidity (−0.762 for mortality), while the correlation between 2006 O/E and change in O/E 2005 to 2007 was −0.013 for morbidity (−0.125 for mortality). Thus, 2-year improvement was

TABLE 2. Coefficient Estimates and Odds Ratios for 2005 Risk Adjustment Models

	Morbidity*						Mortality*					
NSQIP Variable (History of . . .)	Coefficient	Odds Ratio	SE	*P*	OR: 95% CI		Coefficient	Odds Ratio	SE	*P*	OR: 95% CI	
Demographics												
Age	0.010	1.010	0.001	<0.001	1.007	1.012	0.043	1.044	0.004	<0.001	1.036	1.052
Male	0.046	1.047	0.041	0.241	0.970	1.129	0.186	1.205	0.121	0.064	0.989	1.467
Race white	0.092	1.096	0.046	0.030	1.009	1.191	−0.053	0.948	0.107	0.636	0.761	1.182
Preoperative Risk Factors												
Ascites	0.581	1.788	0.214	<0.001	1.415	2.260	1.166	3.208	0.543	<0.001	2.303	4.469
Chemotherapy	0.423	1.527	0.195	0.001	1.190	1.960	0.599	1.821	0.512	0.033	1.049	3.161
Coma	0.156	1.168	0.542	0.737	0.470	2.902	1.066	2.904	1.268	0.015	1.234	6.834
Current pneumonia	0.208	1.232	0.216	0.235	0.873	1.737	0.062	1.063	0.240	0.786	0.683	1.656
Cerebrovascular accident with deficit	−0.055	0.947	0.115	0.651	0.747	1.200	−0.297	0.743	0.193	0.253	0.447	1.236
Cerebrovascular accident without deficit	−0.026	0.974	0.112	0.820	0.778	1.220	0.025	1.025	0.233	0.914	0.657	1.600
Diabetes (any)	0.105	1.110	0.056	0.038	1.006	1.226	0.119	1.126	0.131	0.308	0.896	1.416
Dialysis	0.295	1.344	0.144	0.006	1.089	1.658	0.898	2.455	0.423	<0.001	1.752	3.440
DNR status	−0.621	0.537	0.085	<0.001	0.394	0.733	0.620	1.859	0.403	0.004	1.215	2.844
Dyspnea (any)	0.278	1.320	0.070	<0.001	1.190	1.464	0.645	1.906	0.211	<0.001	1.534	2.367
>2 drinks alcohol/wk	−0.019	0.981	0.096	0.847	0.810	1.188	−0.016	0.984	0.231	0.945	0.621	1.558
Hemiplegia	0.106	1.112	0.188	0.529	0.798	1.550	0.471	1.602	0.519	0.145	0.850	3.022
Angina	0.379	1.461	0.230	0.016	1.073	1.988	0.134	1.143	0.329	0.642	0.650	2.009
Congestive heart failure	0.232	1.261	0.166	0.079	0.974	1.632	0.036	1.037	0.212	0.859	0.695	1.547
Chronic obstructive pulmonary disease	0.290	1.337	0.097	<0.001	1.160	1.540	0.216	1.241	0.171	0.117	0.947	1.626
Myocardial infarction	0.093	1.098	0.173	0.553	0.807	1.494	0.234	1.263	0.306	0.334	0.786	2.031
Peripheral vascular disease	0.053	1.054	0.088	0.530	0.894	1.242	−0.012	0.989	0.169	0.946	0.708	1.381
Transient ischemic attack	−0.114	0.892	0.090	0.257	0.732	1.087	−0.108	0.898	0.200	0.627	0.580	1.389
Hypertension (requiring medication)	0.103	1.109	0.049	0.019	1.017	1.208	0.056	1.057	0.123	0.633	0.842	1.328
Impaired sensorium	0.305	1.356	0.183	0.024	1.040	1.768	0.544	1.723	0.292	0.001	1.235	2.403
Sepsis	0.660	1.936	0.115	<0.001	1.722	2.176	0.776	2.173	0.267	<0.001	1.707	2.765
Previous cardiac surgery	−0.028	0.972	0.067	0.679	0.850	1.112	−0.166	0.847	0.124	0.259	0.635	1.130
Previous percutaneous coronary intervention	−0.022	0.978	0.074	0.767	0.843	1.134	−0.198	0.821	0.137	0.238	0.591	1.139
Radiotherapy	0.071	1.074	0.152	0.615	0.814	1.416	0.176	1.193	0.420	0.616	0.599	2.377
Renal failure	0.447	1.563	0.272	0.010	1.111	2.200	0.831	2.295	0.497	<0.001	1.501	3.510
Rest pain	0.435	1.545	0.160	<0.001	1.261	1.893	0.324	1.383	0.297	0.132	0.907	2.107
Tobacco use	0.245	1.278	0.056	<0.001	1.172	1.393	0.105	1.110	0.136	0.392	0.874	1.411
Steroid use	0.320	1.377	0.104	<0.001	1.188	1.597	0.170	1.185	0.200	0.314	0.851	1.651
Transfusion	0.661	1.937	0.372	0.001	1.330	2.821	0.096	1.101	0.275	0.700	0.674	1.798
Ventilator dependent	1.298	3.662	0.539	<0.001	2.744	4.887	0.996	2.709	0.465	<0.001	1.935	3.791
Wound infection	0.332	1.394	0.108	<0.001	1.198	1.622	−0.028	0.972	0.153	0.858	0.715	1.322
Weight loss (10% in 6 mo)	0.321	1.378	0.109	<0.001	1.181	1.609	0.711	2.037	0.322	<0.001	1.494	2.776
Operative Information												
Wound class clean	−0.374	0.688	0.046	<0.001	0.603	0.785	−0.432	0.649	0.104	0.007	0.474	0.889
ASA class I–II	−0.640	0.527	0.024	<0.001	0.481	0.577	−1.598	0.202	0.042	<0.001	0.135	0.302
Emergency case	0.391	1.479	0.079	<0.001	1.331	1.643	0.993	2.700	0.308	<0.001	2.159	3.376
Vascular case	−0.324	0.723	0.093	0.011	0.563	0.930	0.428	1.534	0.437	0.134	0.877	2.683
Work RVU's	0.059	1.061	0.002	<0.001	1.056	1.065	0.037	1.037	0.007	<0.001	1.025	1.050
Regression constant	−3.004						−7.651					

	2005 Cases†	2006 Cases†	2007 Cases†	2005 Cases	2006 Cases	2007 Cases
C-statistic	0.805	0.816	0.811	0.927	0.934	0.934
Hosmer-Lemeshow (prob > χ^2)	16.4 (0.037)	40.4 (0.000)	667.4 (0.000)	21.5 (0.006)	20 (0.010)	97.1 (0.000)

*Regression also included procedure family codes, as indicated in Table 1, which are not shown.
†Statistics calculated for applying single adjustment algorithm to case populations defined in Table 1.

TABLE 3. Changes in Hospital Performance (O/E*)

A. 118 Institutions Present 2006–2007	Change 2006–2007* Morbidity	Change 2006–2007* Mortality
Mean change in O/E	**−0.1137**	**−0.1740**
P (mean not zero)	**<0.000001**	**<0.0001**
t test (paired) of O/E's for years	4 × 10e-14	<0.0001
Standard error	0.0139	0.0423
95% confidence upper limit for mean	−0.0861	−0.0902
95% confidence lower limit for mean	−0.1412	−0.2579
Median	−0.0906	−0.0926
Minimum	−0.5135	−2.9783
Maximum	0.2173	1.0229
Volume weighted mean	**−0.1126**	**−0.1631**
% Institutions improved	**82%**	**66%**

B. 35 Institutions Present 2005–2007	Change 2005–2006 Morbidity	Change 2005–2006 Mortality	Change 2006–2007 Morbidity	Change 2006–2007 Mortality	Change 2005–2007 Morbidity	Change 2005–2007 Mortality
Mean change in O/E	**−0.0399**	**0.1236**	**−0.1225**	**−0.2286**	**−0.1624**	**−0.1050**
P (mean not zero)	**>0.3**	**>0.05**	**<0.00001**	**<0.001**	**<0.001**	**0.1450**
t test (paired) of O/E's for years	0.3160	0.0550	2.3×10e-6	0.0004	0.0060	0.1400
Standard error	0.0392	0.0621	0.0216	0.0582	0.0432	0.0695
95% confidence upper limit for mean	0.0398	0.2499	−0.0785	−0.1104	−0.0746	0.0362
95% confidence lower limit for mean	−0.1196	−0.0026	−0.1664	−0.3469	−0.2501	−0.2462
Median	−0.0581	0.1055	−0.1088	−0.1870	−0.1713	−0.0358
Minimum	−0.5393	−0.6872	−0.4966	−1.0683	−0.9136	−1.7555
Maximum	0.7627	0.8177	0.1051	0.3936	0.7067	0.5831
Volume weighted mean	**−0.0538**	**0.1224**	**−0.1149**	**−0.1977**	**−0.1555**	**−0.0664**
% Institutions improved	**54%**	**34%**	**89%**	**80%**	**89%**	**57%**

C. 83 Institutions Present Only 2006–2007	Change 2006–2007 Morbidity	Change 2006–2007 Mortality
Mean change in O/E	**−0.1099**	**−0.1510**
P (mean not zero)	**<1 × 10e-7**	**<0.01**
t test (paired) of O/E's for years	1.8 × 10E-8	0.0074
Standard error	0.0176	0.0550
95% confidence upper limit for mean	−0.0749	−0.0417
95% confidence lower limit for mean	−0.1450	−0.2604
Median	−0.0848	−0.0530
Minimum	−0.5135	−2.9783
Maximum	0.2173	1.0229
Volume weighted mean	**−0.1115**	**−0.1472**
% Institutions improved	**80%**	**60%**

*Change in performance calculated as: Recent year O/E minus prior year O/E; negative quantity represents improvement.
Bold indicates highlighted findings.

more strongly associated with the initial performance assessment than the "midterm" assessment.

Net Improvement: 83 Institutions Present Only From 2006 to 2007

Of the 118 institutions present 2006 to 2007 (discussed above), there were 83 institutions that were present only in this period and not prior. Results for this group (Table 3) confirm that overall changes seen remained significant improvement for both morbidity (80% institutions improved, mean change: −0.110) and mortality (60% improving, mean change: −0.151), even when longer participating hospitals were removed.

Comment on Volume Weighting

Reporting results only on an institutional basis has potential to mislead. Weighting each institution's results by their volume effectively converts the analysis to "patient-based." In general, volume weighting made the change in O/E seen in any period slightly less negative. Stated otherwise, there appeared to be slightly less improvement across the entire treated population with volume weighting. However, no changes between weighted and unweighted were significant. For the 10 comparisons in Table 3, after volume weighting, 7 of 10 had slightly smaller improvements (average: 0.014 O/E unit change), while the remaining 3 had negligibly larger changes (average: 0.006 O/E unit change). In no case, after volume

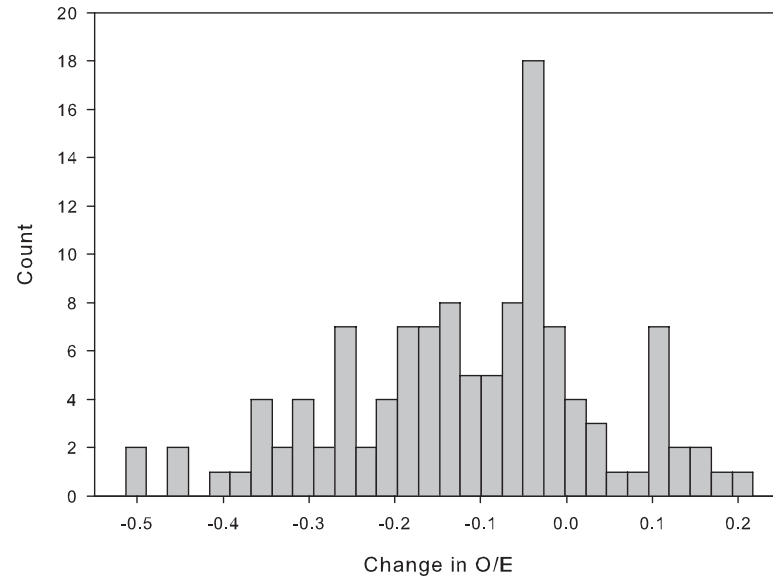

FIGURE 1. Change in morbidity O/E for 118 institutions, 2006–2007. A negative number for change represents reduction of O/E, thus improvement. The mean change is negative (improvement), and the population is asymmetrically skewed toward improvement.

weighting, did a significant finding become nonsignificant, nor did a nonsignificant finding become significant.

Outlier Status and Performance Changes: 35 Institutions Present From 2005 to 2007

An important performance metric for the efficacy of the NSQIP as a QI program is whether bad outliers in particular improve over time. These are the institutions that need improvement the most, and that receive the strongest information signal. Therefore, outliers were first examined for whether they improved at all, and then whether their outlier status changed. Table 4 shows 2005 outlier status versus subsequent improvement in each year. In general, in 2007, 89% of all institutions improved on morbidity, 80% on mortality. Bad outliers participated in this improvement. Table 5 displays the evolution of outlier status over time for these institutions, with risk adjustment held constant, revealing several important conclusions. First, the number of good outliers is dramatically increasing over time. By 2007, according to 2005 standards, 54% of institutions were good outliers for morbidity, and 26% for mortality. In addition, the number of bad outliers falls strongly for morbidity, and after an initial bump, also falls for mortality. Thus, the improvement seen in Table 4 is not insubstantial, as impressive changes in status show up in Table 5.

Outlier Status and Performance Changes: 118 Institutions Present From 2006 to 2007

Examining outlier status for this group yields similar insights to those above. Of the 118 institutions present in both 2006 and 2007, 20 were bad outliers for morbidity at end of 2006: 20 of 20 improved in 2007, and 13/20 were no longer outliers. Fifteen had been called bad outliers for mortality at end of 2006: 14 of 15 improved in 2007 and only 1 remained a bad outlier. At end of 2007, (per 2005 adjustment) there were only 7 bad outliers for morbidity and 2 for mortality; but 70 good outliers for morbidity and 26 for mortality. Thus, across all scenarios the numbers of good outliers increased and bad outliers decreased impressively from 2006 to 2007.

This institutional group was large enough to also directly study the magnitudes of changes for outliers, and the results confirm substantial improvements. Bad morbidity outliers in 2006 had a mean change of −0.261 (improvement), bad mortality outliers −0.816 (improvement). Examining outliers also helps argue against regression to the mean as an explanation for these effects (which we will return to below). For morbidity, good outliers had a mean change −0.027 (also improvement), while for mortality, good outliers had a mean change of 0.248 worsening (a smaller magnitude than the change for bad outliers). Thus, these changes for good and bad outliers do not appear to describe a symmetric (on zero), normal distribution, as might be expected for random changes in repeated measures for a group. Furthermore, bad outliers improved more (magnitude) than the remainder of the population as a whole (t statistic for population means: 6 × 10E-6 for morbidity, 0.001 for mortality), and more than good outliers (t statistic 5 × 10E-8 for morbidity, 3 × 10E-5 for mortality). These statistics reinforce the revelations of correlations already presented.

Outlier Status: Relative Risk of Improvement by Bad Outliers

To build on the information above for magnitude of change by bad outliers, we examined the likelihood for bad outliers to improve versus other institutions. For the 35 institutions present 2005 to 2007, outlier status in 2005 was related to subsequent improvement in 2006, 2007, or aggregated over 2 years. For the 118 institutions present (2006–2007), outlier status in 2006 was compared with improvement in 2007. Bad outliers had a pooled relative risk of improvement of 1.87 compared with good outliers (95% CI: 1.06–3.33; χ^2 relative risk differs from 1 = 4.60, df = 1, P = 0.032). Compared with all institutions except bad outliers, bad outliers had a pooled relative risk for improvement of 1.27 (95% CI: 1.10–1.45; χ^2 relative risk differs from 1 = 11.22, df = 1, P = 0.0008). Our 95% confidence intervals identified 8 bad morbidity outliers for 2005 in this set, while the actual 2005 semiannual report (with 99% intervals) had identified 5 (all captured by our approach). We flagged 3 bad outliers for mortality, while the NSQIP identified 2 (both captured by our approach). Therefore, our analyses might slightly underestimate the true effect of outlier status, since our intervals slightly dilute the historical designation.

Examining morbidity alone, pooled relative risk estimates remained >1 but were not significant (1.34 vs. good outliers, P = 0.15; 1.16 vs. all, P = 0.10). For mortality alone, results were strongly significant. Bad outliers had a relative risk for improvement of 5.76 versus good outliers (95% CI: 2.00–16.56; χ^2 = 10.56, df =

TABLE 4. Initial Outlier Status Versus Subsequent Improvement: 35 Institutions 2005–2007

Morbidity			Morbidity			Mortality			Mortality		
Good Outliers 2005	Improved in 2006	Improved in 2007	Bad Outliers 2005	Improved in 2006	Improved in 2007	Good Outliers 2005	Improved in 2006	Improved in 2007	Bad Outliers 2005	Improved in 2006	Improved in 2007
		x	x		x			x			x
	x	x	x	x	x			x			x
	x		x	x							
		x			x		x			x	
		x	x		x			x			x
	x	x		x	x	x		x			x
		x			x			x			x
	x	x		x	x		x	x	x	x	x
x		x			x			x			x
x		x			x		x	x		x	x
	x	x		x	x		x	x		x	x
x	x	x		x	x			x			x
x		x			x		x	x		x	x
	x		x	x			x			x	
		x			x			x	x		x
		x			x			x			x
	x	x	x	x	x		x	x	x	x	x
x								x			x
x		x			x		x	x		x	x
	x	x		x	x			x			x
	x	x	x	x	x		x	x		x	x
		x			x		x			x	
	x	x	x	x	x			x			x
x		x			x			x			x
	x	x		x	x		x			x	
	x	x		x	x						
	x	x		x	x			x			x
	x	x		x	x			x			x
		x			x	x		x			x
	x			x				x			x
	x	x		x	x		x			x	
		x			x			x			x
	x	x		x	x			x			x
	x	x		x	x			x			x
x		x			x			x			x

Institutions in random order but order preserved throughout.
Improvements are for each year. 2006 and 2007 combined is not depicted.

1, $P = 0.0012$), and a relative risk of 1.47 versus all other institutions (95% CI: 1.24–1.73; $\chi^2 = 19.92$, $df = 1$, $P < 0.0001$).

Outlier Status–Summary of Conclusions

1. The number of bad outliers decreased over time, and the number of good outliers increased over time. Both effects were dramatic. According to 2005 adjustment, by 2007, 70 of 118 hospitals were good outliers for morbidity, 26 of 118 for mortality.
2. Bad outliers were more likely to improve (relative risk) than good outliers or all others, and had statistically significantly larger mean changes than good outliers or all others, but changes were not symmetric, arguing against regression.
3. Bad outliers for mortality had larger improvements and more likelihood of improvement than for morbidity, although morbidity improvements were more uniformly distributed across the entire population.

Holding the Patient Population Constant

We investigated the alternative perspective of passing the same patient population through the same hospitals each year. For the 35 institutions present during 2005–2007, we used only their patients for each year (2005: 33,124 cases; 2006: 49,531 cases; 2007: 48,719 cases) and generated the best risk adjustment model for each year. Then, the 2005 patient population was processed through each of these models to yield the following estimates. The institution-based estimated morbidity for these patients fell from 12.4% in 2005 to 12.16% in 2006 and to 10.76% in 2007. The institution-based estimated mortality fell from 2.02% to 1.79% and

TABLE 5. Outlier Status Over Time: 35 Institutions 2005–2007

Morbidity Good Outliers			Morbidity Bad Outliers			Mortality Good Outliers			Mortality Bad Outliers		
2005	2006	2007	2005	2006	2007	2005	2006	2007	2005	2006	2007
			x	x							
		x	x								
			x								
			x	x	x			x		x	
	x	x				x		x			
		x									
	x	x							x		x
x		x									
x		x									
		x						x			
x	x	x									
x	x	x						x			
			x								
		x							x	x	
										x	
			x						x		
x	x							x			
x	x	x						x			
										x	
			x								
		x		x			x	x			
			x	x	x					x	
x					x						
	x	x					x				
	x	x									
	x	x									
				x		x	x	x			
	x	x					x	x			
				x						x	
		x									
	x	x									
x	x	x									
Number of Institutions That are Good or Bad Outliers											
8	12	19	8	6	3	2	4	9	3	6	1

then 1.49%. Thus, over 2 years there was a relative decrease of 13.2% for morbidity and 26.2% for mortality. Per year this is a compounded reduction of 6.8% morbidity and 14.1% mortality.

Clinical Impact–Complications Avoided

Absolute numbers of adverse outcomes avoided can be estimated by considering each institution's observed and expected complication numbers based on actual volumes accrued. Since these analyses did not require subtracting over 2 time periods, all participating institutions could be analyzed in each year. Table 6 reveals that, relying on the traditional NSQIP outcomes of any morbidity and mortality, substantial numbers of these outcomes appear to be avoided per institution. For any morbidity, all sets and periods show improvement. For mortality, the 35 institution set in 2006 showed a slight worsening, but subsequently improved, and all other numbers reflect improvement.

Furthermore, in 2007, hospitals participating in the NSQIP reported an average of 9000 inpatient surgeries performed and >11,000 outpatient procedures. Since the target annual sample for NSQIP is 1600 to 1800 cases, it is reasonable to assume that the NSQIP represents a sample of roughly 10% to 20% of cases. Obviously, for smaller hospitals the sampling rate is higher, and for larger hospitals lower. For the purpose of estimation factors of 5× and 10× are applied in Table 6, to more closely approximate true numbers.

Table 6 also presents the alternative approach of modeling counts of morbidities, which might correct for underestimation caused by treating any morbidity as an all-or-none phenomenon. Based on this modeling, in 2007, 183 participating hospitals potentially avoided ~9598 complications (mean: 52.5 per hospital, median: 40.6). The distribution of complications avoided per hospital is

TABLE 6. Adverse Events Potentially Avoided

	2006						2007					
	Sample-Based		Scaled Morbidity		Scaled Mortality		Sample-Based		Scaled Morbidity		Scaled Mortality	
Institution Set*	**Morbidity**	**Mortality**	**5×**	**10×**	**5×**	**10×**	**Morbidity**	**Mortality**	**5×**	**10×**	**5×**	**10×**
37 institutions (present 2005–2007)†												
Total no. "any" avoided	165	−39.9					987.3	129.4				
No. "any" avoided per hospital	4.7	−1.1	23.5	47	−5.5	−11	26.7	3.5	133.5	267	17.5	35
121 institutions (present 2006–2007)‡												
Total no. "any" avoided	597.1	17.1					3695	432.5				
No. "any" avoided per hospital	4.9	0.14	24.5	49	0.7	1.4	30.5	3.6	152.5	305	18	36
183 institutions (present 2007)§												
Total no. "any" avoided							4668.5	434.7				
No. "any" avoided per hospital							25.5	2.4	127.5	255	12	24
Total no. counts avoided							9598					
mean no. counts avoided per hospital							52.4		262	524		
P (difference not zero, for counts)							0.002					
95% CI for mean counts avoided, lower							41.6					
95% CI for mean counts avoided, upper							63.2					
Median no. counts avoided per hospital							40.6					

*As these analyses do not require subtracting 1 period from another, additional numbers of institutions could be included in each period.
†2005, 37 institutions: 33,712 cases, mean: 911, SD: 612. For 2006, 35 institutions. For 2007, 37 institutions.
‡2006, 121 institutions: 113,768 cases, mean: 940, SD: 498.
§2007, 183 institutions: 195,402 cases, mean: 1068, SD: 546.

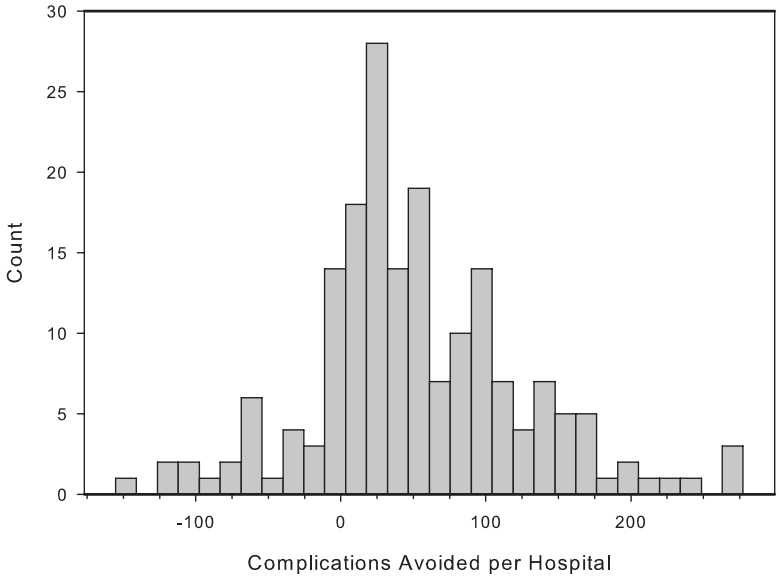

FIGURE 2. Distribution of number of complications avoided by hospitals, 2007. The x-axis represents the number of complications avoided (Expected minus Observed) so a positive number is favorable.

displayed in Figure 2. Again, these numbers likely represent only a fraction (1/5–1/10?) of the total cases performed, so improvements are multiplied. This "counts" approach is not necessary for mortality, which is an all-or-none event. Overall, Table 6 indicates that in 2007 each institution may have avoided 200 to 500 complications, and 12 to 36 deaths.

Changes in Case Mix (Patient Risk Factors)

Returning to the populations described in Table 1, Table 7 shows that the number of comorbidities coded per patient increased slightly each year: 2.11% in 2006; 1.51% in 2007. However, under the constant 2005 risk adjustment algorithm, estimated risk for the populations increased more substantially by year: 3% and 12% for morbidity, 12% and 10% for mortality (Table 7). This raises the question of whether "optimization" or "inflation" of codes is occurring. If there were purposeful manipulation, one might expect this to be most prominent in institutions receiving bad evaluations. However, comparison of O/E ratio point estimates at the end of each time period, versus subsequent change in risk in the next time period, does not reveal significantly biased relationships (Fig. 3).

TABLE 7. Change in Risk Factors and Estimated Risk by Year Using 2005 Adjustment Algorithm

	2005	2006	2007
Institutions	35	118	118
Cases	33,124	112,069	155,058
Risk factors summary			
Comorbidities/pt	1.556	1.589	1.613
Standard deviation	1.747	1.794	1.779
Growth in no. comorbidities/pt–1 yr		**2.11%**	**1.51%**
Growth in no. comorbidities/pt–2 yr			**3.66%**
30-d morbidity			
Average estimated risk	12.41%	12.81%	14.34%
% change in risk–1 yr		**3.22%**	**11.94%**
% change in risk–2 yr			**15.55%**
30-d mortality			
Average estimated risk	1.66%	1.86%	2.05%
% change in risk–1 yr		**12.05%**	**10.22%**
% change in risk–2 yr			**23.49%**

Bold indicates highlighted findings.

DISCUSSION

Evaluating whether quality improvement occurs longitudinally in health care is a challenging task. Methodological issues to consider include regression to the mean, the choice of patient-based versus institution-based analyses, and changes in coding behavior ("code optimization" or "code inflation"). Regarding whether regression to the mean explains the findings in this present study- the answer appears to be "No." First, the net change was toward improvement, not neutral as might be expected for symmetric regression. The distribution of changes was not normally distributed and centered on zero: more institutions improved than worsened over time. In addition, the number of good outliers was always increasing, and bad outliers were stable or decreasing. Both the magnitude of improvement and numbers of institutions improving showed a trend increasing over time, not regressing in each period. Also, there were differential relative risks by outlier status. While regression to the mean must operate to some degree, the magnitude of its effect will be determined by the proportion of assessment due to random "noise." In this case the proportion of that effect does not overwhelmingly explain the observations.

Next, it is reasonable to ask whether our results were weakened by issues of patient-based versus institution-based analysis. Both perspectives are informative, serving different purposes. It is rational to want to know whether, overall, all patients received better care over time, but also rational to want to know how each institution's performance changed. Our analysis was mainly institution-based, but we also presented volume-weighted results, equivalent to a patient-basis. Weighting yielded only very slightly different morbidity, mortality, and improvement rates, and in no case did volume weighting change conclusions regarding direction of change (improvement) or significance.

Third, changes in coding behavior (referred to generously as code optimization or skeptically as code inflation) are potentially always an issue, whether dealing with administrative codes or abstracted clinical data. It is therefore a concern that changes in coding behavior could potentially be contributing to our observed results. Although patient "risk" does appear to be increasing over time, it is increasing more than the number of comorbidities coded, implying fairly sophisticated manipulation if it were purposeful. "Under-coding" of outcomes could potentially result in a similar effect on overall improvement, but our analyses of unadjusted rates of events argue against under-coding. In addition, there is no apparent bias to changes in risk over time, according to institutional O/E, as might be expected for purposeful manipulation. Still, determining the precise role of coding issues, for this and other studies, requires further study.

Moving beyond methodological issues in the arena of clinical surgical quality evaluation and improvement, earlier studies of the VA and private sector NSQIPs have concluded that the programs led to aggregate improvement over time.[1,3] Despite this, whether and to what degree individual institutions improve has remained a pressing question. In this unique, multiple-period assessment of hospital performance, we find convincing indicators of improvement over time. For 118 hospitals from 2006 to 2007 there was a reduction of morbidity by ~0.11 O/E unit (compare with an O/E of 1.0) and mortality by 0.17. For 35 hospitals from 2005 to 2007, there were reductions of 0.16 (morbidity) and 0.11 (mortality). The numbers of hospitals achieving improvement continually climbed, as did the magnitude of improvement. In addition, the number of good outliers rose dramatically, while the number of bad outliers fell. Bad outliers improved at least as much, and probably more, than other hospitals. By 2007, according to 2005 standards, fully 59% of hospitals were "outstanding" performers for morbidity; 22% for mortality. Furthermore, a constant patient population's outcomes improved 6.8% annually (morbidity) and 14.1% annually (mortality). Estimation of counts indicated that tracking any morbidity as an end point may underestimate improvement.

In the current climate of healthcare reform, a recurring issue is: what is surgical quality, and how can it be measured and improved? To date, much effort has been focused on identifying process measures that have links to outcomes. In surgery, this has been the basis for the Surgical Care Improvement Project.[11] While this program has been based on the best evidence in the published data, some limitations do exist, including findings that hospitals with high Surgical Care Improvement Project adherence can still have poor risk adjusted outcomes (personal communication, Dr. Angela Ingraham, ACS). Other analyses of process measures have similarly found little to no correlation between process adherence and risk adjusted outcomes. The NSQIPs focus on clinical outcomes, though not necessarily to the exclusion of process, seems appropriate, and has been gaining support.

In surgery, therefore, relying directly on risk-adjusted outcomes appears to be a sound option for evaluating quality. In the spirit of Codman, collecting and benchmarking "end results" has broad appeal to surgeons, often more so than process measures. However, the issue of data source is important to recognize. While administrative claims data are inexpensive and comparatively easy to acquire, their use for outcomes, particularly complications, has been shown to be limited. While clinically derived data are more difficult to acquire, studies indicate that clinical data abstraction can more completely identify appropriate outcomes. For instance, one comparison of administrative data and NSQIP outcomes demonstrated that the former missed approximately 40% of complications.[12] Such failings are highly relevant when trying to build engagement in surgical QI. The NSQIPs reliance on clinical data is an expense, but also a true strength.

An important question worthy of further discussion is how providers achieve QI once risk adjusted outcomes are known. In this study, we have shown that hospitals participating in ACS-NSQIP appear to improve over time–but how is this achieved? Prior studies have reported that visits to outlier hospitals can identify good versus bad outliers, but actually achieving improvement has been elusive.[13]

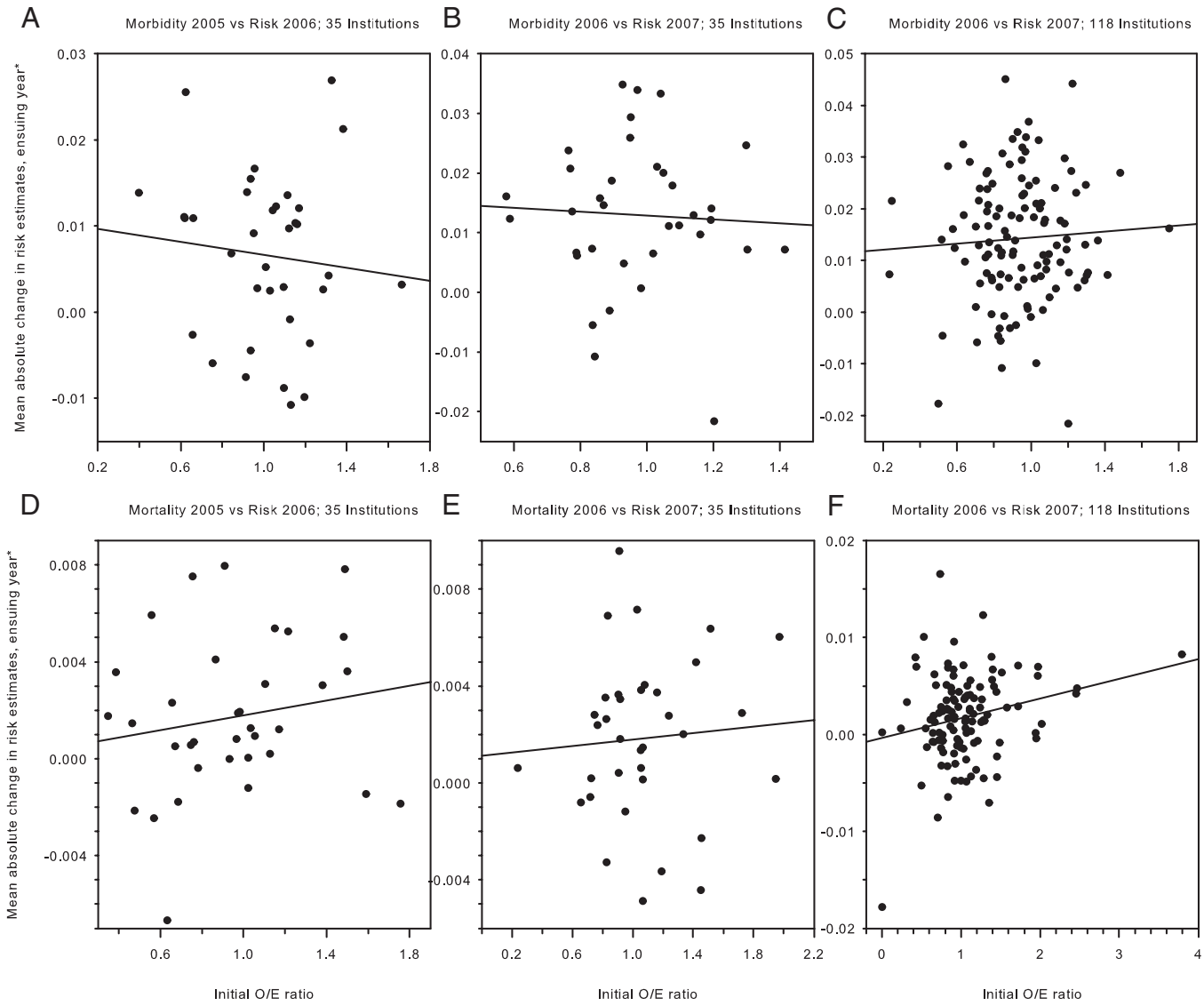

FIGURE 3. Mean absolute change in risk estimate versus starting O/E. The x-axis represents the O/E ratio for the institution at the end of 1 reporting period, while the y-axis represents the mean change in risk estimate in the institution's patient population in the ensuing period. $R^2 < 0.05$ for all. The lack of a strong relationship does not support purposeful code manipulation on the basis of bad O/E.

Currently, the ACS-NSQIP has developed a number of tools to help hospitals improve, including best practices guidelines, case studies of hospitals improving, and rapid data feedback for monitoring progress. Despite these resources, we suspect that not all NSQIP hospitals actively use their information in QI efforts. Not knowing whether institutions act on their data, and not knowing how they act, is a limitation of this study, but it is one that potentially leads to underestimation of the true effect of performance information. The potential for performance improvement may be even greater than we think.

It is important to acknowledge that different measures (in the case of NSQIP–morbidity and mortality) pose different measurement challenges, have different incidences, and can have disparate implications for QI, not least because they can respond differently to QI interventions. For instance, while mortality might seem like a more definite and reliable end point than various morbidities, its utility is often limited by its rarity. While mortality has been a reliable outcome and quality measure for certain procedures, such as CABG, morbidity or selected complications are generally more common and may therefore be a more appropriate as measures for most surgical procedures. In this work, we found significant differences in improvement in morbidity and mortality over time—with morbidity appearing to change more quickly, but mortality potentially changing to a larger magnitude. For morbidity, in all periods the mean changes showed improvement, reaching significance in 2007 and over 2005–2007. In addition, the number of institutions improving was always greater than 50%. For mortality, there appeared to be an initial lackluster response, followed in 2007 by a mean change that was negative and significant, with >50% of institutions improving.

Several particularly interesting findings warrant reemphasis. As stated, over the 2-year period, the number of institutions improving increased cumulatively, and improvements were of greater

magnitude over time. Optimistically, this might reflect that as the ACS-NSQIP has worked with more institutions, the efficacy and success of the program have grown. Another finding of interest is that institutions with worse O/E ratios appeared to change more over time, and had a higher likelihood for change. While institutions across the spectrum of O/E ratios improve, bad performers may be capitalizing on this information. Furthermore, translating improvement into events potentially avoided, not only confirmed all findings of improvement, but also reflected very large clinical impacts in terms of patients affected. In fact, modeling counts confirmed that significant numbers of morbidities appear to be avoided over time, and revealed that the any morbidity end point may underestimate the improvement effect. These results have implications for other "all-or-none" measures. Finally, whether we held risk adjustment constant, or held the patient population risk profile constant, the analyses still indicated that improvement occurred: multiple perspectives were confirmatory.

There are potential limitations of this study. We acknowledge the following:

1. The NSQIP is a self-selected set of programs, which might have propensity for taking on QI. Success might not be universally generalizable.
2. We cannot isolate other trends/programs and influences (local/national), affecting the quality of surgical care over time. NSQIP is not the only factor operating.
3. These data are based on sampling. While the program is changing this approach, current results might still have sampling induced error.
4. These results have a limited focus: general and vascular surgery. Again, generalizability could be questioned.
5. No risk adjustment is perfect. For instance, perhaps there is incomplete control for procedure-specific effects. The program is evolving toward a procedure focus to ameliorate this concern. For these analyses, our inclusion of procedure family indicators improves upon past adjustment.
6. These analyses are based on "opportunities for improvement," which are over more than one period, but are still not long term time trends. A separate analysis of longer-term institutions is currently underway (initial results also indicate improvement).
7. Our adjustment algorithm was not multilevel. In other work, we have found only small differential effects for hierarchical versus nonhierarchical adjustments. In this work, the number and set of institutions evaluated under a constant adjustment varied over time, and institutions were compared primarily to themselves for 2 periods. The importance of a multilevel model in this application is diminished.

In summary, NSQIP institutions appear to be improving morbidity and mortality over time. Multiple perspectives of hospital groupings and time periods support this conclusion. Improvement is seen across the spectrum of all hospitals, with bad outliers in particular possibly capitalizing on performance information. Per institution and year, potentially hundreds of complications and dozens of deaths appear to be avoided. The contribution of changes in risk to these observations requires further study. Despite some limitations on these findings, institutional improvement in the ACS-NSQIP appears to be significant and reflect substantial clinical impact.

REFERENCES

1. Khuri S. The NSQIP: a new frontier in surgery. *Surgery*. 2005;138:837–843.
2. Khuri S, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg*. 1998;228:491–507.
3. Khuri S, Henderson W, Daley J, et al. Successful Implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the Private Sector: The Patient Safety in Surgery Study. *Ann Surg*. 2008;248:329–336.
4. Daley J, Forbes M, Young G, et al. Validating risk-adjusted surgical outcomes: site visit assessment of process and structure. *J Am Coll Surg*. 1997;185:341–351.
5. Daley J, Khuri S, Henderson W, et al. Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. *J Am Coll Surg*. 1997;185:328–340.
6. Khuri S, Daley J, Henderson W, et al. The National Veterans Administration Surgical Risk Study: risk adjustment for the comparative assessment of the quality of surgical care. *J Am Coll Surg*. 1995;180:519–531.
7. Khuri S, Daley J, Henderson W, et al. Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. *J Am Coll Surg*. 1997;185:315–327.
8. Khuri S, Henderson W, Daley J, et al. The patient safety in surgery study: background, study design, and patient populations. *J Am Coll Surg*. 2007;204:1089–1102.
9. American College of Surgeons: ACS-NSQIP (2008). Data collection form, data definitions, protocols, qualifications/training/ auditing of coordinators, and additional database information. Available at: https://acsnsqip.org/main/resources_downloads.asp; https://acsnsqip.org/main/program_data_collection.asp; https://acsnsqip.org/main/get_started_requirements.asp https://acsnsqip.org/main/resources_faqs.asp; https://acsnsqip.org/documents_section/documents_chapter4.pdf. Accessed June 11, 2009.
10. Hall B, Hsiao E, Majercik S, et al. The impact of surgeon specialization on patient mortality: examination of a continuous Herfindahl-Hirschman index. *Ann Surg*. 2009;249:717–718.
11. Centers for Medicare and Medicaid Services (CMS), Surgical Care Improvement Project (SCIP). Available at: http://www.qualitynet.org/dcs/ContentServer?c=MQParents&pagename=Medqic/Content/ParentShellTemplate&cid=1122904930422&parentName=Topic. Accessed April 11, 2009.
12. Steinberg SM, Popa MR, Michalek JA, et al. Comparison of risk adjustment methodologies in surgical quality improvement. *Surgery*. 2008;144:662–667; discussion 662–667.
13. Campbell DA Jr, Henderson WG, Englesbe MJ, et al. Surgical site infection prevention: the importance of operative duration and blood transfusion–results of the first American College of Surgeons-National Surgical Quality Improvement Program Best Practices Initiative. *J Am Coll Surg*. 2008;207:810–820.

Discussions

DR. RALPH G. DEPALMA (WASHINGTON, DISTRICT OF COLUMBIA): Given the superiority of detailed chart review by expert nurse reviewers over administrative data, along with expense considerations for these personnel, what future role do you see for the clinical nurse reviewers in gathering 100% of the quality data? Given the demonstrable differences in O/E modeling for specific procedures (for example, abdominal aortic aneurysm, solid organ transplant, colon surgery, or bariatric surgery), what value does your group find in looking at using actual morbidity and mortality figures to evaluate total facility performance in terms of appropriate and timely presurgical referral and care? There may be a little bit of heresy here with regard to risk adjustment, but we should recognize that failure to rescue is the most common contributor to surgical mortality and morbidity.

Your choice to select confidence limits of 0.95 for both mortality and morbidity differs considerably from mortality and morbidity criteria used by the VA NSQIP program. The VA deliberately set a mortality level of 0.90, a very low radar screen, for defining mortality outliers. Morbidity was set at 0.99. What threshold would you consider designating for "bad outliers?" Might these parameters be changed in the future? What kind of effector arms or oversight structures do you envision in the future to devise site visits and best practice guidance for facilities (bad outliers) falling below

the 0.95% bar? What statistical or time related disciplines would be used to structure site visits should you decide to use active oversight? Finally, to asses comparable performance analyses, would you comment on the feasibility of using a group of all hospitals participating in a 3-year study and modeling all patients at those hospitals for all years combined? Will you consider such an approach?

DR. BRUCE L. HALL (ST. LOUIS, MISSOURI): Regarding detailed chart review versus administrative coding, we in the ACS-NSQIP feel convinced that the expense and effort of detailed chart review is worthwhile. A number of studies, notably one of Ohio State (Steinberg et al. *Surgery.* 2008;144:662–667), showed that certain administratively based algorithms for identifying problems like this fail to identify important events, and so we absolutely feel there is a role for that expense. Must it be performed by a clinical nurse reviewer? That has been our policy so far, but we recently changed that policy to a well-trained and examined, and qualified reviewer, mainly because of a shortage of nurses. We feel that if the reviewer is trained and audited and qualified, and largely independent of the surgeons, then this is an acceptable and robust approach. Do we favor collection of 100% of procedures? Yes. As you may know, we recommended moving the ACS-NSQIP forward with collection of 100% of certain procedure families, or procedure buckets. That effort was led by Dr. Birkmeyer and me on behalf of our Modeling and Evaluation Committee. We feel that this provides advantages both in modeling and in our ability to expand what we can say about the performance of institutions and, in particular, about the performance of certain individuals. One of the biggest, most impressive improvements we can make to risk adjustment modeling is to model limited procedure families. This is a prime basis for that change. Whether we reach all the way to 100% collection, or just much closer, the changes will improve the program. You point out that we provide different models for different procedures in standard NSQIP reporting now and you asked whether there is a utility for using all actual morbidity and mortality rates for institutional performance evaluation for referrals. I do not think I can answer that question definitively at present. That is, as you know a controversial political issue. As you know, the VA has undertaken what appears to be a very logical rationalization of facility capabilities and procedures that may be a shining example for all of us to examine, but I am not sure that I can say that this absolutely should be the basis for referrals at the present time. Regional referrals are a complex issue that involves surgeons knowing their capabilities within the context and capabilities of the hospital and its staff. More work is needed in this area.

You are correct, in this current work we used 95% confidence intervals, and as you point out, the NSQIP traditionally employed 90% for mortality and 99% for morbidity. In fact, our method leads to a slight underestimation of performance improvement by us in this manuscript, in comparison to the existing traditional NSQIP method. For the sake of consistency and standard statistical criteria, we settled on 95% intervals for this analysis because of the way we had to manipulate the analysis over time. But again, we feel that this underestimated improvement. Which criteria are most appropriate for determination of outliers? I do not believe there is 1 answer. We can rely on standard statistics tradition, or we can set new thresholds for the purpose at hand. In the future in the NSQIP I think we will present not only confidence intervals but also percentile based performance reviews, which institutions report to us as useful despite a lack of statistical significance in many situations. By contrast, if the issue is performance-based pay and not just QI, then I think the strictest criteria are necessary. As far as effector mechanisms, the ACS-NSQIP maintains several categories of mechanisms, and is always striving to develop additional ones. Currently we support mechanisms such as: dissemination of best practices and published reviews; case studies; support of institutional collaboratives (regional and procedure specific); and monthly surgeon champion calls. As far as exactly how we will approach site visits for outliers, both in practical terms and in regard to statistics, again I do not think there is 1 answer. First, we relied in the past and will continue to rely on multiple assessments over time, like the VA system does. However, I might argue that the ACS-NSQIP is designed to be more of a "collaborative" improvement program than a program based on certifications and policing. Currently, we have no punitive functions that we apply to institutions. Effective QI probably requires each hospital actively performing its own evaluations and re-evaluations over time. We see our purpose as helping all institutions improve, perhaps poorly functioning institutions most of all. So I believe the answer to this question is in evolution. You asked whether it is feasible to model this question of performance improvement over time using all hospitals all years. Yes, actually, it is not only feasible but might be preferable to examine longer time trends in this way. As I believe you know, Tracy Schifftner Smith in the VA took that approach to looking at probably more than 10 years of performance in the VA, and she models all institutions in 1 group using years as indicator variables. That is a very sound approach. We did not apply that in this study because of the shorter period evaluated. In separate work, we are doing this for our institutions that have been present for 6 years or more. It is a smaller set for us than for the VA, but those preliminary results also confirm that institutions are improving. That is perhaps the preferred approach to identifying long-term time effects, and I agree with your remarks along those lines.

DR. HENRY A. PITT (INDIANAPOLIS, INDIANA): Certainly, at Indiana University Hospital ACS-NSQIP is helping us improve our outcomes. In its present form, this program works very well for academic and teaching hospitals in urban centers for general surgeons and vascular surgeons. However, ACS-NSQIP does not work very well for surgical specialists at those hospitals, or for community surgeons in smaller hospitals where the programs is too expensive, or in the academic medical centers for the endocrine surgeons who want to keep track of laryngeal nerve problems, or for pancreatic surgeons who want to know about pancreatic fistulas. I know that the ACS-NSQIP leadership is working on moving the program forward. How do we get ACSNSQIP to work for all surgeons in all hospitals?

DR. BRUCE L. HALL (ST. LOUIS, MISSOURI): As you point out, there is a bias in NSQIP toward academic and urban institutions and a bias of focus toward general and vascular surgery, although other specialties are represented in a multispecialty model. As you know, the Modeling and Evaluation Committee of the NSQIP recommended that we move forward with a much tighter procedure specific-focus defining categories or buckets of closely related procedures and modeling them on their own. This will enable us to include new specific independent risk factors for various procedures, but also new specific outcomes or dependent factors for each procedure. This will enable us to tell the colorectal surgeon more about their anastomotic leaks, to tell the endocrine surgeons more about their hypocalcemia. This is an ongoing concerted effort involving champions from various surgical societies around the country providing advice and helping to create these new foci and these new sets of procedures, and we are very optimistic that this will improve the value of the program for the populations you describe. We are very sensitive to the issue of rural and community hospitals. Again, we are biased against them in terms of participation and presence currently, but we made a concerted effort to approach those hospitals and to get them involved in large numbers in the program. In fact, the College sponsored a grant to approach more than

1200 critical access hospitals, get them into the program, provide them the kinds of information they need, and also do so at a cost that will be commensurate with their resources. We hope that within the next year or 2 we will have evidence that we achieved those aims.

DR. DAVID R. FLUM (SEATTLE, WASHINGTON): It is an act of faith that surveillance and benchmarking will improve surgical performance, and the exciting part is trying to tie that to an act of science. You spent a lot of time looking at regression to the mean, and I would caution you in your interpretation of it. I would like to hear your response to this 1 component. In your example of what you would call "bad outliers" —a term that we probably should avoid, movement in mortality rates is all over the place, up, and down. That looks like regression to the mean. I wonder if perhaps regression to the mean can happen in hospitals that are outliers in a negative fashion in a different way than it happens in hospitals that are outliers in a good fashion. In other words, for hospitals that have more leaks 1 year than another or more deaths 1 year than another, a regression that bounces around will regress to the mean over time. What are you seeing in the long haul with bad outliers as they move to the mortality rankings?

DR. BRUCE L. HALL (ST. LOUIS, MISSOURI): I agree with your concern. We tried to comment carefully on the fact that bad and good outliers behave differently. In addition, not only individual changes for outliers, but also the progressive accumulating improvements seen, argue against regression alone. I agree there is bound to be noise and there must be a contribution of regression. We just do not feel that that overwhelmingly explains these results. We will certainly investigate these questions further over time.