

# Toward Robust Information: Data Quality and Inter-Rater Reliability in the American College of Surgeons National Surgical Quality Improvement Program

Mira Shiloach, MS, Stanley K Frencher Jr, MD, MPH, Janet E Steeger, RN, BSN, Katherine S Rowell, MHA, MS, Kristine Bartzokis, RN, BSN, MHA, Majed G Tomeh, MS, MBA, Karen E Richards, BS, Clifford Y Ko, MD, MS, MSHS, FACS, Bruce L Hall, MD, PhD, MBA, FACS

- 
- BACKGROUND:** Data used for evaluating quality of medical care need to be of high reliability to ensure valid quality assessment and benchmarking. The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) has continually emphasized the collection of highly reliable clinical data through its program infrastructure.
- STUDY DESIGN:** We provide a detailed description of the various mechanisms used in ACS NSQIP to assure collection of high quality data, including training of data collectors (surgical clinical reviewers) and ongoing audits of data reliability. For the 2005 through 2008 calendar years, inter-rater reliability was calculated overall and for individual variables using percentages of agreement between the data collector and the auditor. Variables with > 5% disagreement are flagged for educational efforts to improve accurate collection. Cohen's kappa was estimated for selected variables from the 2007 audit year.
- RESULTS:** Inter-rater reliability audits show that overall disagreement rates on variables have fallen from 3.15% in 2005 (the first year of public enrollment in ACS NSQIP) to 1.56% in 2008. In addition, disagreement levels for individual variables have continually improved, with 26 individual variables demonstrating > 5% disagreement in 2005, to only 2 such variables in 2008. Estimated kappa values suggest substantial or almost perfect agreement for most variables.
- CONCLUSIONS:** The ACS NSQIP has implemented training and audit procedures for its hospital participants that are highly effective in collecting robust data. Audit results show that data have been reliable since the program's inception and that reliability has improved every year. (*J Am Coll Surg* 2010;210:6–16. © 2010 by the American College of Surgeons)
- 

Quality assessment and quality improvement are critical goals in the US health care system and in the profession of surgery. There are now many different approaches to assess-

ing quality, but all are founded on some type of data—clinical, administrative, or both—and are either prospectively or retrospectively obtained. Not surprisingly, the reliability of the source data underlying many of these quality assessment efforts varies tremendously, which has implications for the trustworthiness of the ultimate assessments obtained.

The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) has relied from its start on robust clinical data, purposefully collected by trained and audited personnel, based on rigorous definitions. ACS NSQIP supports the ongoing acquisition of high quality data through several mechanisms, including rigorous data field definitions with ongoing review, detailed training, and support of dedicated data collection personnel (the surgical clinical reviewer [SCR]), and ongoing evaluation of the reliability of data collected. We de-

**Disclosure Information:** Nothing to disclose.

Received June 30, 2009; Revised September 16, 2009; Accepted September 22, 2009.

From the Division of Research and Optimal Patient Care, American College of Surgeons, Chicago, IL (Shiloach, Richards, Ko); the Robert Wood Johnson Clinical Scholars Program, UCLA/RAND, Los Angeles, CA (Frencher); QCMetrix, Inc, Waltham, MA (Steeger, Rowell, Bartzokis, Tomeh); the Department of Surgery, University of California, Los Angeles, and VA Greater Los Angeles Healthcare System, Los Angeles, CA (Ko); and the Department of Surgery, John Cochran Veterans Affairs Medical Center, the Center for Health Policy, and the Olin Business School, and the Department of Surgery, School of Medicine, Washington University in St Louis, St Louis, MO (Hall). Correspondence address: Mira Shiloach, Division of Research and Optimal Patient Care, American College of Surgeons, 633 N Saint Clair St, Chicago, IL 60611.

### Abbreviations and Acronyms

ACS NSQIP	= American College of Surgeons National Quality Improvement Program
IRR	= inter-rater reliability
SCR	= surgical clinical reviewer
STS	= Society of Thoracic Surgeons
VA	= Department of Veterans Affairs

scribe the mechanisms applied within the ACS NSQIP to promote acquisition of reliable data, and provide evidence that the reliability is very high.

## METHODS

The general methods of the NSQIP have been described in detail.<sup>1-4</sup> Briefly, NSQIP collects clinical data on patients undergoing major surgical procedures, including preoperative demographics and comorbidities, intraoperative information, and morbidity and mortality outcomes for 30 days after operation.

The ACS NSQIP evolved from the US Department of Veterans Affairs (VA) NSQIP, started in 1994 as a method of providing reliable, risk-adjusted surgical outcomes data that could be used to assess surgical quality within and among VA medical centers.<sup>5-9</sup> To determine if the methodology of the VA program could be applied to the private sector, the VA NSQIP collaborated with the American College of Surgeons on a grant-funded study involving 18 private-sector hospital participants from 2001 to 2004 (Agency for Healthcare Research and Quality grant #5U18HS11913-03, "Reporting System to Improve Patient Safety in Surgery").<sup>4</sup> After successful implementation of the study, the ACS NSQIP became an open subscription program at the end of 2004, so 2005 was the first full year of open subscription in the ACS NSQIP. For this report, we present results of inter-rater reliability (IRR) assessments for the calendar years 2005 through 2008. The current organizational structure of ACS NSQIP is such that administration and management of the program are conducted by the American College of Surgeons; technical and clinical issues (including data audits) are managed in conjunction with a team of technical and content expert coordinators at QCMetrix, Inc.

The program collects roughly 135 variables for each accrued patient case in the following general categories: surgical profile, preoperative risk factors, preoperative laboratory data, operative data, and postoperative occurrences, as indicated in [Table 1](#). One of the main goals of the ACS NSQIP data collection is to create statistical models that describe the observed-to-expected ratio for mortality and morbidity outcomes for each participating hospital, adjust-

ing for differences in patient risk factors between hospitals. Using these models, each hospital can compare its outcomes in a reliable manner with those from other medical centers in the program. High quality data are imperative for this process.

The program effects acquisition of high quality clinical data through the following mechanisms:

- Dedicated reviewer for data collection
- Initial reviewer training
- Ongoing online reviewer training and examination
- Continuous reviewer support system
- Creation and continual review of rigorous data definitions
- Dilemma resolution
- Checks for data integrity
- Inter-rater reliability audits
- ACS NSQIP expert coordinator team

### Dedicated reviewer for data collection

The ACS NSQIP has maintained the approach of having a dedicated SCR<sup>10</sup> to collect data. Advantages to this approach include having one or more persons who can be specifically trained and monitored in a uniform manner across all participating sites. In addition, distractions to the SCR's functions are minimized. The reviewer maintains some separation and independence from the surgeons in the program (to limit bias in data collection), but usually has working relationships with them, which is necessary for the success of the program in obtaining the required pieces of information relevant to ACS NSQIP. A disadvantage of the dedicated reviewer is that this can be costly in terms of salary and support.

### Initial reviewer training

Hospitals enrolling in the ACS NSQIP send their SCR to a 2-day training session at ACS headquarters. The training session relays basic information about the program and its structure, the data collection process including the random sampling methods, locating relevant medical information, and detailed review of the program variables and definitions. In addition, the SCRs are trained on the use of the ACS NSQIP software for data entry and practice collecting and entering case data. After in-person training, the SCRs complete a posttest to ensure their understanding of the material. A score of 90% is required to pass the posttest.

### Ongoing online reviewer training and examination

In addition to the in-person training, each SCR also completes six online training modules as a part of the certification process. Modules 1 and 2 present an overview of the

**Table 1.** Inter-Rater Reliability for Audited Variables

Variable	Disagreement between raters, %			
	2005	2006	2007	2008
Overall percent disagreement	3.15	2.26	1.99	1.56
Audited variables, n	38,978	96,990	144,690	140,132
Individual variables				
Surgical profile				
Year of birth	0.26	0.00	0.37	0.23
Gender	1.06	0.87	0.37	0.61
Race	0.53	2.92	1.03	1.06
CPT code	1.85	6.12	1.99	2.27
Status (inpatient/outpatient)	1.06	0.00	0.88	1.06
Transfer	3.17	3.79	2.14	1.59
Hospital admit date	1.32	0.29	0.22	0.53
Operation date	0.79	0.00	0.22	0.30
Anesthesia technique	1.06	1.17	0.96	1.59
Subspecialty	0.53	0.29	0.74	0.45
Preoperative risk factors				
Height	6.08	3.50	1.55	1.44
Weight	4.23	1.75	0.88	1.21
Diabetes	3.70	3.21	3.17	2.95
Current smoker	6.61	2.33	2.43	1.36
Pack year cigarette history	19.05	10.79	7.22	3.48
Alcohol use	0.79	1.46	0.96	1.89
Dyspnea	8.73	5.25	5.45	3.25
Do not resuscitate status	5.29	0.29	0.96	0.68
Functional health status - before current illness	9.52	6.41	2.87	1.97
Functional health status - before operation	11.38	7.00	4.93	3.40
Ventilator dependent	0.53	0.29	0.52	0.38
History of severe COPD	2.12	3.50	3.53	3.40
Current pneumonia	1.06	0.87	0.88	0.68
Ascites	1.32	1.46	0.88	0.98
Esophageal varices	0.00	0.58	0.15	0.15
Congestive heart failure	3.70	2.04	1.47	1.21
History of myocardial infarction	1.85	1.46	0.52	0.53
Previous PTCA	2.38	3.50	2.50	3.40
Previous cardiac surgery	0.79	1.17	1.91	1.29
History of angina	2.38	0.29	1.10	1.74
Hypertension requiring medication	4.76	4.66	3.76	2.50
History of revascularization/amputation for PVD	2.38	2.62	1.99	2.12
Rest pain/gangrene	4.50	2.92	2.95	1.36
Acute renal failure	1.59	0.87	1.40	1.51
Current dialysis	0.53	0.29	0.44	0.15
Impaired sensorium	3.17	2.04	0.81	0.45
Coma	0.00	0.00	0.00	0.08
Hemiplegia/hemiparesis	1.32	1.75	1.55	1.51
History of transient ischemic attack	2.38	0.58	0.96	1.59
CVA/stroke with neurologic deficit	2.65	1.17	1.40	0.76
CVA/stroke with no neurologic deficit	2.91	2.33	1.40	1.44
Tumor involving central nervous system	0.00	0.29	0.15	0.08

(continued)

**Table 1.** Continued

Variable	Disagreement between raters, %			
	2005	2006	2007	2008
Paraplegia/paraparesis	0.53	0.58	0.29	0.38
Quadriplegia/quadruparesis	0.00	0.00	0.15	0.15
Disseminated cancer	1.32	1.46	1.33	1.36
Open wound	6.35	4.66	2.65	2.12
Steroid use	1.32	2.04	0.96	1.66
> 10% loss of body weight	2.38	2.33	1.33	1.44
Bleeding disorders	10.05	4.66	5.67	5.22
Transfusions > 4 U	0.79	0.29	0.37	0.23
Chemotherapy	0.53	0.29	0.37	0.38
Radiotherapy	0.79	0.29	0.37	0.30
Sepsis (SIRS/sepsis/septic shock)	7.67	6.71	6.85	4.46
Pregnant	0.00	0.00	0.00	0.00
Earlier operation within 30 d	0.00	0.00	1.33	1.44
Preoperative laboratory values				
Sodium	8.47	4.96	5.15	3.63
Blood urea nitrogen	7.67	5.54	5.67	4.31
Creatinine	7.14	4.96	5.15	2.57
Albumin	5.56	4.37	5.74	3.78
Total bilirubin	5.82	5.25	5.23	3.03
Aspartate transaminase	5.29	4.96	5.96	3.18
Alkaline phosphatase	5.82	5.25	6.92	4.01
WBC	9.52	2.62	5.45	3.4
Hematocrit	9.52	4.08	5.60	3.78
Platelets	9.52	4.08	5.23	4.31
Partial thromboplastin time	5.82	5.54	5.15	4.84
International normalized ratio	6.61	4.08	5.30	3.48
Prothrombin time	7.41	4.96	5.74	3.86
Operative data				
Other CPT codes	7.14	9.04	3.17	2.87
Concurrent CPT codes	1.59	1.46	1.18	0.91
Emergency case	2.12	1.75	3.31	2.57
Wound class	6.61	7.29	11.63	10.89
American Society of Anesthesiologists class	2.65	2.92	1.77	1.82
Transfusion (intraoperative)	2.65	1.17	1.55	0.76
Operation start time	0.53	0.58	0.59	0.30
Operation finish time	1.32	0.29	0.37	0.30
Death during operation	0.00	0.00	0.00	0.00
Cardiac arrest requiring CPR	0.00	0.29	0.07	0.00
Unplanned intubation (intraoperative)	0.53	0.29	0.07	0.23
Myocardial infarction (intraoperative)	0.00	0.00	0.15	0.00
Postoperative occurrences				
Superficial incisional surgical site infection	4.76	5.54	1.62	2.04
Deep incisional surgical site infection	2.65	2.33	1.33	1.29
Organ/space surgical site infection	1.06	1.17	1.10	1.21
Wound disruption	0.79	1.46	0.96	1.66
Pneumonia	1.06	2.04	1.91	0.98
Unplanned intubation	2.65	1.46	0.66	0.76
Pulmonary embolism	0.26	0.29	0.15	0.08

(continued)

**Table 1.** Continued

Variable	Disagreement between raters, %			
	2005	2006	2007	2008
On ventilator > 48 h	4.76	0.58	0.96	0.68
Progressive renal insufficiency	0.79	0.58	0.66	0.61
Acute renal failure	0.53	0.00	0.74	0.38
Urinary tract infection	2.12	1.75	1.62	1.21
Cerebral vascular accident	0.00	0.00	0.22	0.08
Coma > 24 h	0.26	0.58	0.15	0.08
Peripheral nerve injury	0.53	0.00	0.07	0.00
Cardiac arrest requiring CPR	1.06	1.17	0.29	0.23
Myocardial infarction	0.79	0.58	0.07	0.53
Bleeding > 4 U packed red blood cells	1.06	0.87	0.74	0.53
Graft/prosthesis/flap failure	0.26	0.58	0.15	0.30
Deep vein thrombosis/thrombophlebitis	0.79	0.00	0.29	0.08
Systemic sepsis	15.08	13.70	7.88	4.01
Discharge data				
Hospital discharge date	0.53	0.00	0.59	0.53
Postoperative ICD-9 code	3.70	2.92	0.59	0.23
Return to operating room	1.06	0.58	0.66	1.06
Death within 30 d	0.26	0.00	0.00	0.08
Death > 30 d	0.26	0.29	0.29	0.08
Date of death	0.26	0.00	0.00	0.08

PTCA, percutaneous transluminal coronary angioplasty; PVD, peripheral vascular disease; SIRS, systemic inflammatory response syndrome.

ACS NSQIP program, processes, and statistical analysis and are completed before attending the in-person training. Modules 3 and 4 present the preoperative, intraoperative, and postoperative definitions using a case scenario format. These modules require the SCR to review cases and complete the data collection form as if they were reviewing a real case. Modules 5 and 6 provide the SCR with additional detail on variables that have caused confusion or resulted in high levels of disagreement between the SCR and coordinators during on-site audits. These two modules also require the use of comprehensive case studies and completion of the data collection form. The last four modules are completed at specific times over the course of the first 6 months of participation in the program.

SCRs must score at least 80% on each module as a requirement for passing. In addition, the SCR is awarded 11.25 continuing education units for completing the set of six modules. The modules offer an email option on each screen that the SCR can use to send comments or questions to the training staff. Review of the comments has allowed for immediate clarification of SCR issues and stimulates ongoing improvements and updates to the module.

### Continuous SCR support system

The SCR's work is continually supported through conference calls, online resources, and the ACS NSQIP annual

conference. These support systems are integral to the reviewer's continuing education, understanding, and development within the SCR role.

Conference calls for SCR are held periodically with expert coordinators from QCMetrix. Calls may be held to discuss a specific topic or may be open to questions and discussion from the SCR. Online support is provided through several tools that assist SCR in determining the correct application of definitions. These tools include Data Definition Lookup, Decision Trees that help reviewers apply criteria to determine whether a variable is present, and a Frequently Asked Questions (FAQs) section, which can be used to check whether a definition query has previously been answered. In addition, SCR may submit any program-related questions through the online Ask a Question function. Responses are provided within 2 business days. Finally, a weekly case study is posted online—the case study tests the SCR's definition knowledge by presenting a hypothetical scenario and asks whether a certain definition applies to the scenario.

The ACS NSQIP annual conference is a valuable resource that allows SCR to network, gain knowledge about the global importance of ACS NSQIP to the surgical community, and to learn how the robust data collected provide an integral step to quality improvement in hospitals.

SCR need to learn and retain a large amount of information to complete their daily data collection responsibilities. The ongoing education allows the SCR to continu-

ously access information necessary to ensure accurate and consistent data collection, and to improve their clinical judgment of the many variables. The ACS NSQIP annual conference reinforces the training received and demonstrates the many potential applications for surgical quality improvement using robust ACS NSQIP data.

### **Creation and continual review of rigorous data definitions**

Data definitions have been developed mainly using a consensus-based process among the ACS NSQIP Data Definition Committee members. The committee consists of six surgeons, two SCRs, a senior expert coordinator and the director of expert coordinators from the QCMetrix site, and two biostatisticians from the statistical group responsible for ACS NSQIP data analysis. Clinical definitions from regulatory agencies are adapted for program use wherever possible. For example, the ACS NSQIP uses the Centers for Disease Control (CDC) definition as the basis for identifying pneumonia and surgical site infection. The goal has been to produce definitions that are specific enough to allow the SCR to easily ascertain the information without requiring clinical guesswork. Specific data definitions also ensure that each SCR at every site is collecting the same information, thereby reducing variability. Data definitions are continually evaluated by a focused committee. The committee reviews selected definitions that have been problematic as evidenced by SCR questions and by poor inter-rater reliability. If necessary, the committee updates the definitions to attempt additional clarification. Updated definitions are distributed to the SCR community through the ACS NSQIP secure Website.

### **Dilemma resolution**

The ACS NSQIP procedure for resolving discrepancies in data collected is outlined in the program manual. The procedure for dilemma resolution directs SCRs to first attempt to internally resolve discrepancies between information found on the surgical record and what they believe should be recorded for ACS NSQIP data collection. For example, if the surgical record indicates that the wound classification is “clean” but the SCR can ascertain that the type of operation would obviously result in a “clean-contaminated” wound, then the SCR should attempt to solve this discrepancy with the surgeon performing the procedure. If necessary, the surgeon champion makes the final decision on the discrepancy. In cases in which the surgeon champion cannot resolve the discrepancy, the SCR contacts the expert coordinator at QCMetrix. The expert coordinator then discusses the issue with the surgeon champion to arrive at a decision. In addition, the SCRs can raise data definitions

issues to be discussed by the Data Definitions Committee, if necessary.

### **Checks for data integrity**

The QCMetrix team continuously monitors site performance and data integrity through data analytic reporting. Reports are run to identify aberrances in site performance and data integrity. Automated checks are used to control illogical entries, such as discharge before admission and other inappropriate entries. Site performance measures include aspects of case selection, rates and patterns of case accrual, median days to case transmission, and percentage of cases transmitted with complete 30-day follow-up. Data integrity measures include comparisons against like sites and national benchmarks for percentages of preoperative risk factors and reported postoperative occurrences. When required, the expert coordinators develop corrective action plans to assist the site toward resolution. One example of such an action plan for a site that has a low 30-day follow-up rate would include regular contacts by the expert coordinator with the SCR to monitor progress, regular monitoring of the 30-day follow-up rate, and focused education. If required, the expert coordinator assists the SCR in escalating issues to the manager and surgeon champion to gain their support in removing barriers to improving the 30-day follow-up rate.

### **Inter-rater reliability audits**

The IRR audit is a fundamental tool of ACS NSQIP to assess the quality of the data collected at participating sites. This process involves the review of 12 to 15 charts per institution and time period audited. Charts are selected based on criteria designed to identify potential reporting errors, such as cases with five or more preoperative risk factors and no reported mortality or morbidity, or cases with two or fewer preoperative risk factors and reported mortality or morbidity. Other cases are selected using random sampling. Operating room logs are also audited to ensure correct sampling of cases. Finally, the IRR site visit assesses the SCR’s program processes to determine if he or she is efficiently and effectively collecting data and to make recommendations for improving processes if needed.

The site visitor reviews approximately 106 variables for each case. The disagreement rate between the SCR and the site reviewer is calculated as a percentage using the number of disagreements divided by the total number of variables reviewed. The disagreements are additionally classified by whether the site reviewer identified the variable and the SCR did not, the SCR identified the variable and the site reviewer did not, or whether they both identified the variable but disagreed on the variable category. The site reviewer assesses whether the SCR has any issues with over-

reporting or under-reporting of specific data categories. The SCR and surgeon champion receive a written report that summarizes the visit findings.

The ACS NSQIP Steering Committee has determined that a variable disagreement rate  $> 5\%$  both for individual variables and overall site audits requires corrective action. If any site has a  $> 5\%$  disagreement rate during the audit, the site reviewer will make an assessment of the contributing factors and work collaboratively with the site to develop a site-specific corrective action plan. In addition, the ACS NSQIP Steering Committee reviews these sites' IRR site visit findings and a follow-up IRR site visit is scheduled 3 to 6 months from the initial audit.

Cumulative audit results are also analyzed, and each variable is reviewed for its program-wide disagreement rate. Variables with a  $\geq 5\%$  disagreement rate for all IRR audits are reviewed with all SCRs as part of the education process. On occasion, certain variables are reviewed by the ACS NSQIP Definitions Committee to determine whether or not the variable can be collected more reliably or if the definition requires additional clarification. These variables become the focus of intensive education for all SCRs in the field through the site visit, regularly scheduled conference calls, case studies, and testing.

### ACS NSQIP expert coordinator team

Many of the previously mentioned training, support, and reliability audit activities are administered by the ACS NSQIP Expert Coordinator Team. Traditionally, most of the coordinators have been registered nurses experienced with outcomes data collection. Their training includes completion of the SCR training class, the initial training posttest with a minimum passing score of 95%, and the six online training modules with a passing score of 90%. In addition, each coordinator completes a minimum of three IRR audits as part of their orientation. The third audit is certified by the director of expert coordinators or an experienced ACS NSQIP expert coordinator to assure that the new expert coordinator is properly completing the site review. Finally, expert coordinators are required to abstract ACS NSQIP program variables from a complex case study. The complex case study is completed annually and expert coordinators must have a minimum score of 90% to pass.

### Analysis and Cohen's kappa estimation

This analysis specifically reports results of IRR audits for 2005 through 2008. These data were aggregated from audits and examinations conducted at each individual site for each period. Rates of disagreement were classified as described earlier under "Inter-rater reliability audits." In addition, Cohen's kappa was estimated for selected variables from the 2007 audit year. Kappa could be estimated only

for dichotomous variables because information on the distribution of ratings of variables with multiple categories was not collected. In addition, it was necessary to make assumptions on the distribution of ratings where both reviewers agreed. In such cases, the distributions were assumed to mirror the prevalence of the variable from the 2007 ACS NSQIP data. Calculations were completed using SPSS version 14.

## RESULTS

### Participation and training

The ACS NSQIP was opened to enrollment in 2005 with 16 of the 18 institutions that were original recipients of the Agency for Healthcare Research and Quality grant continuing their participation in the program. Sixty-two hospitals were enrolled by the end of 2005, 150 enrolled by the end of 2006, and 205 enrolled by the end of 2007. Two hundred twenty eight institutions were participating in the program, as of the end of 2008.

Four hundred nineteen SCRs participated in the initial training session, as of December 31, 2008, completed the initial training posttest, and are currently collecting program data or serving as back-up to the primary SCR. Eighty-five percent of the SCRs passed the training posttest on their first attempt with an average score of 96%. The 15% who failed on their first attempt scored an average of 85%, and on their second attempt scored an average of 96%. Trainees must score 80% or above on each of six online training modules, as a part of the requirements for SCRs to collect data. Average scores for the online training modules range between 91% and 95%, as of the end of 2008. The ACS NSQIP Expert Coordination Team average scores on the initial training posttest and six learning modules were 100% and 95%, respectively.

### Inter-rater reliability audits

Figure 1 indicates the numbers of institutions enrolled, audits performed, and charts reviewed during these time periods. The auditing functions of the program have grown in step, as the number of institutions has grown. Until July 2008, the program policy was to audit each hospital within its first year of participation, and then every other year thereafter, unless the hospital failed an audit and required a follow-up audit. The audit policy was subsequently changed such that only new sites or sites with a new SCR were audited, audits every other year were eliminated, and targeted audits will be implemented for sites based on an annual assessment of data. This change in audit policy accounts for the decreased number of charts reviewed seen in Figure 1.

Table 1 provides complete detail on the inter-rater reliability of every variable during the periods examined. Of

**Table 2.** Variables with Inter-Rater Disagreement Greater than 5%

Variables	Disagreement by year, %			
	2005	2006	2007	2008
Preoperative risk factors				
Pack year smoking history	19.05	10.28	7.22	
Functional health status before surgery	11.38	7.08		
Bleeding disorder	10.05	6.68	5.67	5.22
Functional health status before current illness	9.52	6.01		
Dyspnea	8.73		5.45	
Systemic sepsis	7.67	8.41	6.85	
Current smoker	6.61			
Open wound	6.35			
Height	6.08			
Do not resuscitate status	5.29			
Preoperative laboratory data				
Hematocrit	9.52	6.41	5.45	
Platelets	9.52	6.01	5.23	
White blood cell count	9.52	5.34	5.45	
Sodium	8.47	6.14	5.15	
Blood urea nitrogen	7.67	6.41	5.67	
Prothrombin time	7.41	6.01	5.74	
Creatinine	7.14	5.61	5.15	
International normalized ratio	6.61	5.34	5.30	
Partial thromboplastin time	5.82	6.14		
Alkaline phosphatase	5.82	5.07	6.92	
Total bilirubin	5.82		5.23	
Albumin	5.56		5.74	
Serum glutamic-oxaloacetic transaminase	5.29		5.96	
Operative data				
Other CPT codes	7.14	7.88		
Wound class	6.61	8.68	11.60	10.89
CPT code		5.61		
Postoperative occurrences				
Systemic sepsis	15.08	14.15	7.88	
Discharge data				
Postoperative ICD-9 code		5.74		

note, the reliability of data within the program (assessed in this way) has been extremely high from the start (3.15% disagreement in 2005), and even more importantly, has continually improved. Overall disagreements on variables have steadily fallen, from 3.15% in 2005 to 1.56% in 2008, as indicated in the table.

A particular aim of the program has been to continually improve any poor-performing variables. To this end, the program puts special emphasis on evaluating and refining variables with disagreement rates > 5%. Table 2 reports variables that met this threshold over time. Disagreement

**Table 3.** Estimated Cohen's Kappa Values for Selected Variables (2007)

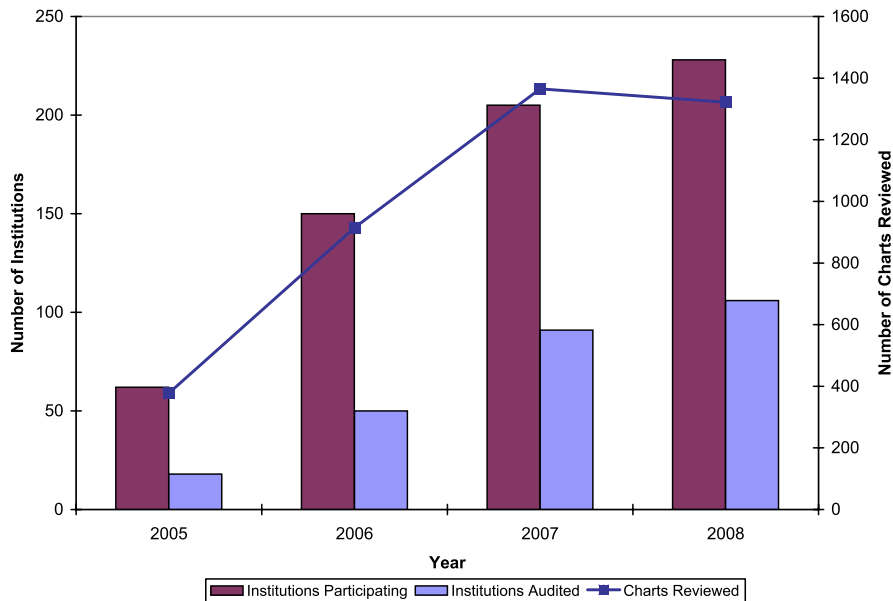
Variable	Estimated kappa	95% CI (Lower limit, Upper limit)
Steroid use	0.86	0.78, 0.93
Current smoker	0.93	0.90, 0.95
Disseminated cancer	0.70	0.57, 0.84
History of peripheral vascular disease	0.77	0.68, 0.85
Radiotherapy	0.73	0.51, 0.95
Rest pain	0.38	0.23, 0.53
Previous percutaneous coronary intervention	0.78	0.71, 0.85
Do not resuscitate status	0.32	0.05, 0.59
Ascites	0.71	0.56, 0.87
History of angina	0.32	0.07, 0.57
Hypertension	0.92	0.90, 0.94
Open wound	0.66	0.56, 0.77
Bleeding disorder	0.38	0.27, 0.49
Weight loss >10%	0.72	0.56, 0.85
Transfusion	0.49	0.08, 0.89
Ventilator dependent	0.50	0.27, 0.73

levels for nearly all problem variables have continually improved, and the number of variables with disagreement > 5% has been reduced over time, from 26 such variables in 2005 to just 2 in 2008. In addition, outlying problem variable disagreement levels also improved over time; the maximum disagreement in 2005 was 19.05%, the current maximum for 2008 is 10.89%.

In general, problematic disagreements have usually resulted from the SCRs' lack of understanding of the variable definition. In the case of wound class, this variable may have a high rate of disagreement because of the operating room nurse incorrectly classifying the wound, and the SCR failing to recognize the misclassification. Additional feedback from the SCRs about wound classification has shown that there was a lack of understanding of bacterial load of the surgical site, where normal flora exists, and in which situations normal microbes become contaminants. In the case of laboratory values, disagreements arose mainly from the SCRs not selecting the laboratory value closest to the date and time of operation (a requirement in the definition). However, extensive improvements have been made in recognizing this particular problem, and disagreement rates for all laboratory values have fallen below 5% for 2008.

### Kappa values

Cohen's kappa statistics were estimated for selected variables in 2007 as described in the Methods section (Table 3).



**Figure 1.** Inter-rater reliability audits over time. American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) participating institutions, audits performed, and charts reviewed for the 2005 through 2008 period.

We emphasize that these are estimates based on the assumptions noted. These estimates reflect inter-rater reliability while accounting for agreements occurring by chance. The closer the estimated kappa value is to 0, the more likely the agreement occurred by chance. Most of the 16 variables calculated had moderate to substantial agreement; only 4 variables had “fair” agreement (defined as kappa value between 0.2 and 0.4). All kappa values were statistically significant ( $p < 0.01$ , data not shown).

## DISCUSSION

The ACS NSQIP takes a number of reinforcing approaches to ensure the acquisition of robust, high quality, reliable data. SCRs receive extensive training and ongoing support to guarantee that all reviewers understand the variable definitions and are collecting consistent data. Audits of participating institutions are used to determine inter-rater reliability and provide an additional educational component for SCRs when disagreements on variable assessments are found. Evaluation of inter-rater reliability shows that the program is successful in achieving outstanding data reliability, and efforts to improve reliability have “been fruitful” since the program’s inception in 2005. The most recent results indicate only 1.56% overall disagreements between reviewers. Additionally, only two individual variables were above the 5% disagreement threshold for 2008. In a continued effort to improve reliability of individual variables, a new learning module was recently released to

improve the SCR’s understanding of complex variables such as wound classification.

The exceptional reliability of these clinical data contributes to the value of the ACS NSQIP. High quality data create confidence among the participants that the data can be trusted and that the hospital participants are receiving accurate information about their surgical outcomes and benchmarking reports.

Comparison of ACS NSQIP’s audit process to data integrity of other large databases is not straightforward because of differing objectives of the data collected for quality improvement and infrequent reporting of audit methods. However, a few national quality improvement initiatives have developed various methodologies for validating collected data, including the use of trained reviewers, peer-review, or both combined.

ACS NSQIP may be most easily compared with the Society for Thoracic Surgeons (STS) National Database, which maintains three outcomes programs for thoracic surgery: adult cardiac, general thoracic and congenital heart surgery.<sup>11,12</sup> Since 1989, this program has captured data on more than 2 million surgical cases from more than 600 hospitals.<sup>13</sup> STS National Database aggregates information collected from 14 designated regions.<sup>11</sup> The number of contributing participants varies in each of the 3 databases, and ranges from 0 to 104 participants in each state.<sup>11</sup>

STS uses a methodology similar to ACS NSQIP to validate the data collected using trained site data managers,

automated data integrity verification systems, external reviewers, and on-site audits of randomly selected charts.<sup>13</sup> STS categorizes its approach according to three levels of data quality assurance: internal, regional, and national.<sup>13</sup> Internally, integrity checks at the point of initial data entry and submission to the national database, in combination with real-time feedback, has led to increases in data validity.<sup>14</sup> The average percent of missing values, as of 2002, on the 28 clinical variables used for mortality risk modeling was 1.6%. Of these, only two variables were missing more than 5% of the time.<sup>13</sup> Regional auditing is exemplified by the Iowa STS region, which receives support for auditing services of its statewide clinical databases through the Iowa Foundation for Medical Care (IFMC).<sup>15</sup> The Iowa Foundation for Medical Care is a private nonprofit organization contracted by the Center for Medicare and Medicaid Services (CMS) as the Quality Improvement Organization (QIO) for Iowa, and as such, conducts annual audits of reported data elements. Similarly to ACS NSQIP, audits examine agreement between STS data abstracted by Iowa Foundation for Medical Care staff and that submitted by hospital staff. Audits particularly emphasize data elements with < 95% agreement in previous audited years. In 2002, agreement on data variables ranged from 80% to 100% with an average of 96.4%.<sup>13</sup> Finally, data integrity is established on a national level by comparing STS data with benchmarks from a national, publicly available database, the Medicare Provider and Analysis Review (MEDPAR). Results of the comparison yielded higher surgical volumes and mortality rates within the National Cardiac Database than in the Medicare Provider and Analysis Review, suggesting completeness of data entry.<sup>13</sup>

Other national databases focus on specific surgical diseases. Denmark, Finland, and Scotland have each developed registries for laparoscopic cholecystectomy, vascular procedures, and hip fracture care. Lessons learned from these early quality improvement programs are reflected in the procedures adopted by ACS NSQIP and the STS National Database. For example, when the Danish Laparoscopic Registry was initially instituted, formal reporting procedures and a validating methodology were not established, leading to wide variability in completeness of reported elements (range 69% to 99%).<sup>16</sup> The Finnish Vascular Registry demonstrated similar rates of incomplete data: 21% in 1994.<sup>17</sup> Even recently established national registries such as the National Lung Cancer Audit in the United Kingdom reported 50% data completeness in its first 2 years.<sup>18</sup> Now in its third year, participation in the National Lung Cancer Audit has reached 75%, with 22,600 lung cancer case submissions.<sup>19</sup> However, nonresponders and incomplete submissions continue to plague

the program. Particularly problematic have been findings among some surgical registries that patients with unreported data tended to have poorer outcomes.<sup>20</sup> ACS NSQIP is aware of the problems that missing data can entail, and has designed its SCR training, data collection and feedback, and audit programs such that incomplete data are virtually nonexistent. Missing data are an issue only in the laboratory values fields, where the test may not have been performed on the patient, and not as a result of the SCR failing to enter the data.

The effect of a dedicated data collector and standardized collection procedures on completeness of data collection is evident in other large disease registries. The Scottish Hip Fracture Audit was started in 1993 to collect data on case-mix procedures, treatment, and outcomes. In 1998, the audit collected 27% of hip fracture cases in Scotland, improving to 63% of cases by 2003.<sup>21</sup> In 2004, the Scottish Hip Fracture Audit built on a national effort to validate and standardize local reporting by consolidating regional databases into an audited national database. Data were collected on site by dedicated coordinators who submitted hospital data to a central team composed of a clinical coordinator, statistician, and data coordinator. The Scottish Hip Fracture Audit collected data using standardized data collection forms and procedures, and implemented dual data entry and electronic validation processes. Subsequently, data completeness has improved to 100% collection of cases in 2008.<sup>22</sup>

Finally, audit procedures of the Healthcare Effectiveness Data and Information Set (HEDIS), conducted by National Center for Quality Assurance (NCQA), can be compared with the ACS NSQIP. The Healthcare Effectiveness Data and Information Set is used by 90% of American health plans to assess 71 measures across 8 domains,<sup>23</sup> and the measures are used to make appreciable comparisons of the performance between health plans. The data collected are validated by certified auditors using a National Center for Quality Assurance process containing both "off-site" and "on-site" components.<sup>2</sup> The off-site process includes verification of compliance with Healthcare Effectiveness Data and Information Set data reporting procedures, survey sample frame validation, selection of core metrics for review, and medical record review. These processes are followed by a site visit, which allows for observation of systems used for data collection, and enhanced education for the data collectors to improve reliability. Published work of the audit results for the years 1997 to 2000 showed that kappa values for inter-rater reliability were overall very high (greater than 0.75) for nearly all measures. Each year, one or two measures showed a need to improve rater reliability

and these measures were targeted for improvement in subsequent years.<sup>24</sup>

Given the potentially transformative nature of quality improvement programs and the data used to guide recommended changes in practice, it is crucial for the data to be as valid as possible. Identifying and investing in auditing mechanisms that reinforce accurate data collection is imperative for this objective. ACS NSQIP has implemented a highly effective training and audit procedure for its hospital participants, with audit results showing very reliable data and improvements in reliability every year.

### Author Contributions

Study conception and design: Shiloach, Steeger, Rowell, Tomeh, Richards, Ko, Hall

Acquisition of data: Steeger, Rowell, Bartzokis, Tomeh, Hall

Analysis and interpretation of data: Shiloach, Bartzokis, Hall

Drafting of manuscript: Shiloach, Frencher, Hall

Critical revision: Shiloach, Frencher, Steeger, Rowell, Bartzokis, Tomeh, Richards, Ko, Hall

### REFERENCES

1. ACS NSQIP participant use file user's guide. Available at: [https://acsnsqip.org/puf/docs/ACS\\_NSQIP\\_Participant\\_User\\_Data\\_File\\_User\\_Guide.pdf](https://acsnsqip.org/puf/docs/ACS_NSQIP_Participant_User_Data_File_User_Guide.pdf). Accessed June 16, 2009.
2. ACS NSQIP program specifics. surgical case inclusion/exclusion overview. Available at: [https://acsnsqip.org/main/program\\_case\\_inclusion\\_exclusion.asp](https://acsnsqip.org/main/program_case_inclusion_exclusion.asp). Accessed June 16, 2009.
3. Fink AS, Campbell DA Jr, Mentzer RM Jr, et al. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. *Ann Surg* 2002;236:344–353; discussion 344–353.
4. Khuri SF, Henderson WG, 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.
5. Daley J, Forbes MG, Young GJ, et al. Validating risk-adjusted surgical outcomes: site visit assessment of process and structure. National VA Surgical Risk Study. *J Am Coll Surg* 1997;185:341–351.
6. Daley J, Khuri SF, 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.
7. Khuri SF, 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.
8. Khuri SF, 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.
9. Khuri SF, 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.
10. ACS NSQIP program specifics. surgical clinical nurse reviewer training. Available at: [https://acsnsqip.org/main/program\\_nurse\\_training.asp](https://acsnsqip.org/main/program_nurse_training.asp). Accessed June 16, 2009.
11. Society of Thoracic Surgeons. US maps of STS National Database participants. Available at: <http://www.sts.org/sections/stsnationaldatabase/new/>. Accessed June 6, 2009.
12. Ferguson TB Jr, Dziuban SW Jr, Edwards FH, et al. The STS National Database: current changes and challenges for the new millennium. *Ann Thorac Surg* 2000;69:680–691.
13. Welke KF, Ferguson TB Jr, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. *Ann Thorac Surg* 2004;77:1137–1139.
14. Ferguson TB Jr, Hammill BG, Peterson ED, et al. A decade of change—risk profiles and outcomes for isolated coronary artery bypass grafting procedures, 1990–1999: a report from the STS National Database Committee and the Duke Clinical Research Institute. *Ann Thorac Surg* 2002;73:480–489.
15. Iowa Foundation for Medical Care (IFMC). IFMC Health Care Quality Programs. Available at: <http://www.ifmc.org/iowa.html>. Accessed June 6, 2009.
16. Dreisler E, Schou L, Adamsen S. Completeness and accuracy of voluntary reporting to a national case registry of laparoscopic cholecystectomy. *Int J Qual Health Care* 2001;13:51–55.
17. Lepantalo M, Salenius JP, Luther M, Ylonen K. Introduction of a population-based vascular registry: validity of data and limitations of registration. The Finnvasc Study Group. *Br J Surg* 1994;81:979–981.
18. Rich AL, Free CM, Beckett P, et al. The National Lung Cancer Audit Database (LUCADA); essential analysis of data quality. *Lung Cancer* 2009;63:S14–S15.
19. UKLCC. United Kingdom Lung Cancer Coalition Briefing 2009. Available at: <http://www.uklcc.org.uk/pdf/NLCA%20-%20briefing%20document%20150509.pdf>. Accessed June 2009.
20. Elfstrom J, Stubberod A, Troeng T. Patients not included in medical audit have a worse outcome than those included. *Int J Qual Health Care* 1996;8:153–157.
21. NHS Scotland. Scottish Hip Fracture Audit national trend analysis report 1998–2004. Available at: [http://www.shfa.scot.nhs.uk/AnnualReport/SHFA\\_Trend\\_Analysis.pdf](http://www.shfa.scot.nhs.uk/AnnualReport/SHFA_Trend_Analysis.pdf). Accessed June 6, 2009.
22. NHS Scotland. Scottish Hip Fracture Audit Report 2008. Available at: [http://www.shfa.scot.nhs.uk/AnnualReport/SHFA\\_Report\\_2008.pdf](http://www.shfa.scot.nhs.uk/AnnualReport/SHFA_Report_2008.pdf). Accessed June 6, 2009.
23. NCQA. HEDIS Compliance Audit: standards, policies and procedures. Washington DC; 2009.
24. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. *Am J Manag Care* 2002;8:787–793.