Original Study

Annals of Surgery Open

OPEN

Within-hospital Temporal Clustering of Postoperative Complications and Implications for Safety Monitoring and Benchmarking Using ACS-NSQIP Data

Mark E. Cohen, PhD,* Yaoming Liu, PhD,* Clifford Y. Ko, MD, MS, MSHS, FACS,*† and
Bruce L. Hall, MD, PhD, MBA, FACS,*‡

Objective: To determine the extent to which within-hospital temporal clustering of postoperative complications is observed in the American College of Surgeons, National Surgical Quality Improvement Program (ACS-NSQIP).

Background: ACS-NSQIP relies on periodic and on-demand reports for quality benchmarking. However, if rapid increases in postoperative complication rates (clusters) are common, other reporting methods might be valuable additions to the program. This article focuses on estimating the incidence of within-hospital temporal clusters.

Methods: ACS-NSQIP data from 1,547,440 patients, in 425 hospitals, over a 2-year period was examined. Hospital-specific Cox proportional hazards regression was used to estimate the incidence of mortality, morbidity, and surgical site infection (SSI) over a 30-day postoperative period, with risk adjustment for patient and procedure and with additional adjustments for linear trend, day-of-week, and season. Clusters were identified using scan statistics, and cluster counts were compared, using unpaired and paired *t* tests, for different levels of adjustment and when randomization of cases across time eliminated all temporal influences.

Results: Temporal clusters were rarely observed. When clustering was adjusted only for patient and procedure risk, an annual average of 0.31, 0.85, and 0.51 clusters were observed per hospital for mortality, morbidity, and SSI, respectively. The number of clusters dropped after adjustment for linear trend, day-of-week, and season (0.31–0.24; $P = 0.012$; 0.85–0.80; $P = 0.034$; and 0.51–0.36; $P < 0.001$; using paired *t* tests) for mortality, morbidity, and SSI, respectively. There was 1 significant difference in the number of clusters when comparing data with all adjustments and after data were randomized (0.24 and 0.25 for mortality; $P = 0.853$; 0.80 and 0.82 for morbidity; $P = 0.529$; and 0.36 and 0.46 [randomized data had more clusters] for SSI; $P = 0.001$; using paired *t* tests) for mortality, morbidity, and SSI, respectively.

Conclusions: Temporal clusters of postoperative complications were rarely observed in ACS-NSQIP data. The described methodology may be useful in assessing clustering in other surgical arenas.

Keywords: benchmarking, cumulative sum, NSQIP, postoperative complications, scan statistics, temporal clustering

INTRODUCTION

Postoperative complications are common and are associated with (1) patient operative risk, mostly captured by patient history, comorbidities, age, gender, and laboratory values, (2) the inherent riskiness of the operative procedure, and (3) a quality metric describing the general performance of the hospital (or provider). With respect to the hospital-performance component, hospitals have different resources, practices, and procedures and employ staff with different levels of training and expertise. These features contribute to a steady-state quality environment that exerts relatively consistent effects on outcomes, though subject to gradual quality improvement or deteriorative trends.

However, there is also potential for rapid performance changes, which might be permanent or transitory, influenced by factors such as sporadic outbreaks of infection (eg, associated with the appearance of antibiotic-resistant bacteria or lapses in infection control practices), abrupt changes in hospital staffing, clinician performance, or treatment protocols, unrecognized equipment malfunctions, and so forth. The American College of Surgeons, National Surgical Quality Improvement Program (ACS-NSQIP) reporting has not traditionally focused on rapid or "real-time" detection of complication events, not on detecting temporal clusters of events within hospitals. This stance results, in part, from a desire for robust risk adjustment and assessment of a fixed and clinically valuable 30-day postoperative follow-up period.

ACS-NSQIP emphasizes quarterly benchmarking reports (semiannual reports and interim semiannual reports), always based on 12 rolling months of data. Because of the time required to clean and model data, analyzed data extend back 6 to 18 months. To provide more current information, ACS-NSQIP also has an "on-demand" application where benchmarked data are available shortly after cases are entered into the registry.

From the *Division of Research and Optimal Patient Care, American College of Surgeons, Chicago, IL; †Department of Surgery, University of California Los Angeles David Geffen School of Medicine and the VA Greater Los Angeles Healthcare System, LA; and ‡Department of Surgery, UC Davis Health System and UC Davis School of Medicine, University of California, Davis, Sacramento, CA.

Disclosure: The authors are the employees of the American College of Surgeons, which owns and directs the ACS-NSQIP, and declare that they have nothing to disclose.

Reprints: Mark E. Cohen, PhD, American College of Surgeons, 633 N. Saint Clair St., 23rd Floor, Chicago, IL 60611-3211. Email: mcohen@facs.org.

Annals of Surgery Open (2024) 3:e483

Received: 5 March 2024; Accepted 19 July 2024

Published online 9 August 2024

DOI: 10.1097/AS9.0000000000000483

On-demand assessments are, like semiannual reports, risk-adjusted and "smoothed" but use data within days of entry into the registry.[1] On-demand users can specify an assessment period of between 1 year and 1 month. However, even this approach might be slow to detect rapid shifts in complication rates because events need to be averaged over some minimal time period.

Cumulative sum (CUSUM) and similar methods have been suggested as one approach for the early detection of changes in event rates. CUSUM's advantage lies in continually evaluating incoming data for a cumulative deviation from preestablished performance bounds. This allows for more immediate identification of a rate shift from expectations. However, CUSUM techniques are not optimally designed to differentiate true (below some probability threshold) from chance rate changes. Rather, calibration and validation processes are typically implemented to tune CUSUM detection thresholds to achieve a desired balance between true and false detections. This limitation is recognized when CUSUM findings are suggested to provide a potential early warning of deteriorating performance rather than a definitive assessment at some prescribed *P* value.[2]

CUSUM's apparent advantage in detecting within-hospital "bursts" or "clusters" of postoperative complications would be valuable to the extent that they exist. However, while there are clear, often well-publicized, epidemiologic events or hospital failures causing rapid increases in events, such as a sterilization machine failure, the contribution of such event clusters to the total postoperative complication burden, as specified in and monitored by programs such as ACS-NSQIP, has not been well-established. While CUSUM has reported cluster-like events in surgery, the issue is confused as some CUSUM implementations are structured to be sensitive to both between-hospital differences in rates as well as to within-hospital shifts. This stems from the "expected" rate (which, when compared with an observed rate, forms a metric for identifying deviation) often being estimated from a model that uses data from many hospitals rather than being hospital-specific.[2–4] Thus, the magnitude of the observed-to-expected difference is influenced by both internal event sequences and the hospital's overall quality. As such, a nominal cluster could be the result of an inconsequential "cluster" being superimposed on aberrant baseline quality.

This study evaluates the temporal clustering of complications within hospitals, independent of comparative hospital quality. There are numerous approaches to detecting temporal clusters. CUSUM is one such technique, but a scan statistic approach might permit a more robust evaluation of detection probability. This well-established and general-purpose method was used to determine the extent to which clusters of postoperative complications occur at rates beyond those expected by chance in the ACS-NSQIP program data. If clusters occur at near chance rates, then there is additional confidence in NSQIP's focus on periodic assessments. If clusters appear commonly, this will motivate additional efforts to use CUSUM or other approaches for the early detection of "out-of-control" processes, for which those methods might have superior sensitivity.

The methodologies described and the evidence provided in this study could guide the design of safety analytics in other surgical areas. Trade-offs between periodic and CUSUM-type assessments are described, and decisions about which provides greater net benefit will depend on the underlying temporal structure and impact of adverse events, which would need to be assessed by area experts.

METHODS

Surgeries from hospitals that had accrued at least 2400 patients into ACS-NSQIP in 2018 or 2019 and had at least 1 case in each month of the 24-month period, were included in the study. More recent data would be influenced by the COVID-19 epidemic, but COVID's influence on perioperative complications was recognized in real time. In this study, interest was in the detection of hospital-level complication clusters under more routine conditions of clusters being driven by local, less ubiquitous causative agents.

We evaluated postoperative clusters of 30-day mortality, morbidity (a composite outcome including surgical site infection [SSI], wound disruption, pneumonia, unplanned intubation, on ventilator >48 hours, postop dialysis, postop renal insufficiency, urinary tract infection, stroke/cerebral vascular accident, cardiac arrest, myocardial infarction, systemic sepsis [sepsis or septic shock], all as defined in ACS-NSQIP), and SSI specifically, where an increased rate could be the result of proverbial "out-of-control" system failures—such as might be related to instrument sterilization or a breakdown in an operating procedure.

For each hospital, for each outcome, for each of 730 days (365 days/year for 2 years), the number of expected events on each postoperative day (from 0 to 30 days) was determined based on the number of patients under study each day and a hospital-specific Cox proportional hazards regression for patients' risk of the event on each day. Five different "empirical" models were constructed: (1) Only patient and procedure risk (defined by outcome- and CPT code-specific linear risk—a proprietary ACS-NSQIP variable derived from a multiyear dataset) were considered (variables were chosen from the standard ACS-NSQIP predictor set using forward selection); (2) patient and procedure risk with the effect of hospital-level linear (quality) trend removed by forcing in a continuous variable with values 1 to 730 depending on the day-of-surgery within the 2-year study period; (3) patient and procedure risk with the effect of operation day-of-week removed by forcing in a 7-level categorical day variable; (4) patient and procedure risk with the effect of season removed by forcing in a continuous variable defined as cos ([360 X (operation day of year/365)]^{o}) —this cosine function yields a maximum season-associated difference between operation dates around December 31/January 1 (around 0^{o} the value is close to 1) and June 30/July 1 (around 180^{o} the value is close to −1); (5) patient and procedure risk with adjustments for trend, day-of-week, and season all added. Thus, these 5 models adjust expected events, in various ways for the effects of patient and procedure, gradual day-by-day trend, day-of-week, and seasonal trends. Clusters detected after adjustment for all these factors suggest the presence of some unknown, out-of-control process.[5]

Scan statistic methodology was used to detect clusters defined here as a number of events within a 30-day window significantly ($P \leq 0.05$) exceeding expectations. Obviously, this is just one possible specification of a temporal range for unusual event concentrations. However, it seems to be a reasonable window width for achieving stability and power. Narrower windows would have neither and wider windows might average out rapid changes.

The 30-day scan window was moved 1 day at a time across each hospital's 2-year observational period. All scan windows were examined for their log-likelihood ratio (this is a measure of how likely the observed number of events within a window is, given the model-derived expected rate). The *P* value for the window with the supremum of the likelihood ratio was estimated by way of Monte Carlo hypothesis testing of that likelihood ranking against rankings from random datasets. If that cluster was significant at $P < 0.05$, cases within that 30-day window were no longer eligible for consideration, and the window with the next largest log-likelihood was evaluated as a potential cluster, and so forth, until no window achieved a preset minimum log-likelihood. This methodology is described in the SaTScan documentation, but we replicated those or equivalent methods using our own programming to provide greater analytic flexibility.[6,7] The primary metric studied was the mean number of hospital clusters within the 2-year period.

While the previously described method ensures that each cluster is significant at $P < 0.05$, it does not provide information as to how many clusters would be expected by chance for any particular dataset, time period, modeling method, and scanning scenario. To provide that point of reference, each hospital's dataset was modified as follows: begin with the understanding that each hospital's patient data were distributed nonuniformly across the 730-day observational period. While fields in the dataset associated with patients' operation dates remained fixed in their temporal location, all other data, for each patient, remained linked together and were randomly reassigned to the fixed operation date fields, with the one exception that the date of event, if there was an event, was recomputed to reflect the appropriate operation date to complication date time interval (Table 1). This randomization will remove all but chance temporal clusters. Thus, only risk adjustment for patient and procedure is applied in the proportional hazards model when analyzing the randomized data. This randomization process was repeated 100 times (for each hospital, for each outcome), and the primary metric was the mean number of hospital clusters within the 2-year period, that now arise when only chance clusters can appear.

Two-year cluster counts were compared using both independent and paired sample *t* tests. It was not clear, a priori, which of these tests would have the greatest power as the paired *t* test's ability to control for variability related to hospital differences might be offset by a reduced sample size, as some pairs would be dropped from the analysis if a model for either pair member failed to converge. Parametric tests were deemed appropriate as sample sizes were sufficiently large to yield normal distributions of sample means despite data skew. Nevertheless, nonparametric tests were conducted and confirmed the results (not reported).

NSQIP hospitals are instructed to sample roughly 1680 cases per year, though at times they collect more or fewer cases for various reasons. Under the general assumption that larger samples increase the likelihood of detecting (at a specified level of statistical significance) effects of smaller magnitude, clusters might be detected more frequently with a 100% sample. Two additional analyses were therefore conducted to assess whether reliance on NSQIP's sampled dataset could bias results against detecting clusters. First, correlations (Spearman because of the anticipated positive skew in hospital sample sizes) between hospital sample size and number of detected clusters were computed. Second, our methods were replicated, as a sensitivity analysis, for the morbidity outcome using a 150-day scan window moving 5 days at a time. While a 150-day window is inconsistent with the intent of detecting a rapid performance shift, this was done to reflect the situation where the window sample size is 5 times larger so that it would structurally represent a hypothetical situation where ACS-NSQIP accrual was conducted at a 5-fold increased sampling rate.

RESULTS

The case selection process yielded 1,547,440 patients from 425 hospitals. Table 2 describes the number of hospitals and the number of clusters within hospitals, over the 2-year timeframe, for the 3 outcomes studied. The number of hospitals studied was often reduced from 425 due to model nonconvergence, particularly for the mortality outcome. While nonconvergence of models can sometimes be remedied by modifying procedure specifications, this was not the case in this study, where nonconvergence resulted primarily from very few or no events for some of these within-hospital models.

The number of clusters observed was very small across all modeling conditions. Focusing on the empirical findings, when temporal clustering was adjusted only for patient and procedure risk, an annual average (see Table 2 for discussion of 2-year versus annualized estimates of cluster detections) of 0.31, 0.85, and 0.51 clusters were observed per hospital for mortality, morbidity, and SSI, respectively. While the effects of the individual adjustments for trend, day-of-week, and season were inconsistent, adjusting the empirical findings for the combined effects of all 3 consistently reduced the mean number of clusters over the 2-year period: 0.31 to 0.24; $P = 0.012$; 0.85 to 0.80; $P = 0.034$; and 0.51 to 0.36; $P < 0.001$; using paired *t* tests. Thus, some modest number of apparent clusters were driven by the combined effects of trend, day-of-week, and season.

There was 1 significant difference between the annual number of clusters in the empirical data with all adjustments and the number of "chance" clusters observed for the randomized data: 0.24 and 0.25 for mortality, $P = 0.853$; 0.80 and 0.82 for morbidity; $P = 0.529$; and 0.36 and 0.46 for SSI; $P = 0.001$. Notably, counts for SSI clusters were greater for the randomization condition, suggesting the chance nature of this finding.

There were 10 significant *P* values in Table 2 involving comparisons between counts observed when there were individual adjustments for trend, day-of-week, or seasonality and other groups (empirical data with patient and procedure adjustment only, empirical data with all adjustments, or randomized data). Consistent directionality in these effects was not observed, and there is not a clear mechanism for the effects. These findings might be related to experiment-wise error rate (multiplicity), given the many tests conducted on correlated data.

It is important to note that hospital cluster counts of 1, 2, 3, … 6, or 7 are similar across empirical and randomized data. This distributional similarity suggests that focusing on the mean number of clusters is not masking a high number of clusters in a small number of poorly performing hospitals within the empirical data. Nonetheless, some practitioners might argue that even small numbers of hospitals or providers with apparent preponderances of clusters would be worth flagging. We would

TABLE 1.

Randomization of 10 Hypothetical Patients Using the Process Described in the Text

Analyzed	Original Data		Randomized Data	
Patient Case	Operation Date	All Other Data	Operation Date	All Other Data
1	Patient A	Patient A	Patient A	Patient G
2	Patient B	Patient B	Patient B	Patient J
3	Patient C	Patient C	Patient C	Patient C
4	Patient D	Patient D	Patient D	Patient A
5	Patient E	Patient E	Patient E	Patient H
6	Patient F	Patient F	Patient F	Patient B
7	Patient G	Patient G	Patient G	Patient I
8	Patient H	Patient H	Patient H	Patient D
9	Patient I	Patient I	Patient I	Patient F
10	Patient J	Patient J	Patient J	Patient E

Operation date remains fixed in the dataset. However, randomization reassigns all other patient features, including procedure, patient demographics and comorbidities, and postoperative events (indexed to days from the operation date) as a set (eg, Patient G's data is randomly assigned to Patient A's operation date). This random reassignment removes other than chance temporal clustering, including effects associated with linear trend, seasonality, and day-of-week. The reported results were for 100 separate randomizations of the data.

TABLE 2.

Total Number of Hospitals where Models Converged, Percentage of Hospitals with the Number of Clusters for Each Outcome, and Mean Number of Clusters

	N of Hospitals (of 425) Where Models Converged	Percent of N of Hospitals With the Indicated Number of Clusters over 2 years (These Patterns are Nearly Identical When Only Paired Data Were Considered)								Mean Number of Clusters (Unpaired data) 2 yr/1 yr	*T* test *P* Values for Mean Number of Clusters		Mean number of Clusters (Paired data) 2 yr/1 yr	Paired *t* test *P* Values for Mean Number of Clusters	
											Base vs Base Plus Adjustment	Random vs Empirical		Base vs Base Plus Adjustment	Random vs Empirical
Mortality		0	1	2	3	4	5	6	7						
Empirical															
Base: adjust for patient & procedure	304	61.2	21.4	12.8	4.3	0.3				0.61/0.31	ref	0.145	0.61/0.31	ref	0.024
Base plus adjust for trend	306	67.7	18.0	10.5	3.6	0.3				0.51/0.25	0.147	0.999	0.51/0.26	0.010	0.594
Base plus adjust for day-of-week	305	57.4	24.3	11.5	5.9	1.0				0.69/0.34	0.307	0.014	0.68/0.34	0.101	0.001
Base plus adjust for seasonality	304	65.5	22.4	8.5	3.6					0.50/0.25	0.114	0.921	0.50/0.25	0.017	0.715
All adjustments	305	69.5	17.4	9.5	3.6					0.47/0.24	0.043	0.571	0.47/0.24	0.012	0.853
Randomization: adjust for patient & procedure	304	66.8	19.4	10.2	3.3	0.3				0.51/0.25		ref	0.51/0.25		ref
Morbidity		0	1	2	3	4	5	6	7						
Empirical															
Base: adjust for patient and procedure	422	24.6	28.2	19.0	15.6	7.6	2.4	2.6		1.71/0.85	ref	0.467	1.71/0.85	ref	0.256
Base plus adjust for trend	422	21.8	30.8	18.7	17.8	7.6	2.4	0.7	0.2	1.70/0.85	0.906	0.527	1.70/0.85	0.812	0.274
Base plus adjust for day-of-week	422	21.6	30.8	16.6	15.2	10.0	5.0	0.5	0.5	1.80/0.90	0.365	0.096	1.80/0.90	0.022	0.009
Base plus adjust for seasonality	423	24.3	27.9	22.1	15.8	6.6	2.1	1.2		1.64/0.82	0.467	0.999	1.64/0.82	0.047	0.907
All adjustments	423	21.5	33.6	21.8	13.9	6.1	2.6	0.5		1.59/0.80	0.237	0.648	1.60/0.80	0.034	0.529
Randomization: adjust for patient & procedure	423	22.5	31.4	21.8	14.0	6.6	2.6	1.0	0.2	1.64/0.82		ref	1.63/0.82		ref
SSI															
Empirical		0	1	2	3	4	5	6	7						
Base: adjust for patient and procedure	414	48.8	21.7	14.3	11.6	1.9	1.5	0.2		1.01/0.51	ref	0.223	1.01/0.51	ref	0.064
Base plus adjust for trend	414	52.4	24.2	13.5	7.0	2.2	0.2	0.5		0.85/0.43	0.048	0.399	0.85/0.43	<0.001	0.253
Base plus adjust for day-of-week	414	50.0	19.8	15.7	10.6	3.1	0.7			0.99/0.50	0.800	0.334	1.00/0.50	0.465	0.148
Base plus adjust for seasonality	415	54.0	20.5	15.7	6.7	1.9	1.2			0.86/0.43	0.061	0.459	0.86/0.43	<0.001	0.341
All adjustments	415	56.6	22.4	14.5	5.0	1.0	0.5			0.73/0.36	<0.001	0.001	0.73/0.36	<0.001	<0.001
Randomization: adjust for patient & procedure	413	46.5	28.6	15.0	7.3	2.2	0.5			0.92/0.46		ref	0.91/0.46		ref

The empirical models incorporate 5 levels of risk adjustment. The randomization models were risk adjusted only for patient and procedure, and counts are based on means across 100 separate randomizations. The number of hospitals considered in the paired *t* test analysis was never more than 2 fewer than listed for the number of hospitals where models converged; this was due to a pair being dropped because the model for one of the pair did not converge. The randomization process will remove effects due to an out-of-control process (defined here as true clusters) as well as effects of trend, day-of-week, and seasonality. The most relevant comparison is between empirical data with all adjustments and the randomization (chance) data. The small number of clusters and the similarity of counts under these 2 conditions, as well as across all conditions within each outcome, is notable. All tests were done on 2-year scan data. 1-year cluster count data, reported below and in the text, is calculated as the 2-year scan results divided by 2 and is adopted as a simplification for expository purposes. However, because of differences in detection opportunities for scan windows moving across 1 year versus 2 years, the 1-year estimate is only a close approximation. Bold values are significant at $P < 0.05$.

not object except that the findings favor chance, meaning the risk of false-positive signals seems quite high.

Using the all adjustments dataset (Table 2), the hospital sample size was not associated (Spearman r was used because of dramatic right skew in hospital sample sizes) with cluster counts for all 3 outcomes (mortality: N = 305; sample size range = 2403–13031; cluster count range = 0–3; r = 0.031; P = 0.595; morbidity: N = 423; sample size range = 2403–13031; cluster count range = 0–6; r = 0.086; P = 0.076; SSI: N = 415; sample size range = 2397–12999; cluster count range = 0–5; r = 0.067; P = 0.172). The sensitivity analysis on the morbidity outcome with a 150-day/5 day-at-time window found an annual average of 0.05 clusters under the randomized condition, compared with 0.82 for the 30-day/1 day-at-time window. This reduction would be driven, in part, by fewer cluster detection opportunities (due to stepping the window 5 days rather than 1 day at a time), as well as better control of short-term random clusters. Even when approximately annualized, the simulated increased accrual did not yield more clusters. Also of importance in this approach, annual cluster counts for empirical data with all adjustments did not exceed the counts observed for randomized data.

DISCUSSION

ACS-NSQIP-style periodic and "on-demand" reports, scan statistics, and techniques such as CUSUM can all be used to assess surgical safety. However, while periodic and on-demand reports are directed toward providing robust risk-adjusted benchmarked comparisons between hospitals, scan statistics and CUSUM are mostly concerned with detecting shifts in event rates within hospitals. This is the case for CUSUM when the expected rate is derived from a hospital-specific model. If the expected rate is estimated from an all-hospital model, CUSUM detections will be driven by both within-hospital changes in rate and comparative quality.

Early detection of out-of-control processes yielding event clusters would clearly be an institutional priority. However, before considering whether ACS-NSQIP and similar programs could be slow to detect rapid, within-hospital changes in event rates, it is important to ascertain whether such events occur. The scan statistic findings indicate that the number of clusters of postoperative complications observed in the empirical data does not appear to be greater than those expected by chance. In addition, the presence of some nominal clusters seems to be influenced by a gradual trend, day-of-week, or seasonal effects. While these might represent important clinical features worthy of investigation, they do not suggest the presence of an out-of-control process.

The scan statistic findings were similar whether the analysis was on all data (unpaired t tests) or restricted to paired data, though paired t tests exhibited greater power. One limitation of this study was data loss due to model nonconvergence. This was an issue for mortality, where as few as 303 models converged out of 425 models attempted, but not an issue for morbidity, where at least 422 models converged, and not for SSI, where at least 413 models converged. However, as nonconvergence would likely be associated with models for hospitals with low event rates, this loss of data would likely result in an overestimation of cluster counts with respect to the original set of 425 hospitals.

While the present scan statistic findings suggest that rapid shifts in event rate (or clusters) are rare, this conclusion conflicts with reports based on other detection methods, such as CUSUM. Higher CUSUM cluster detection rates in other work might be attributed to 3 factors. First, by adjustment of detection thresholds, CUSUM can be made more or less sensitive to changes in rate and, when set to high sensitivity, many detections might be false positives. Second, some implementations of CUSUM derive the expected rate from an all-hospital model. As a result, detections could be influenced by small, inconsequential event clusters superimposed on a general quality differential. Finally, most implementations of CUSUM do not appear to adjust for trend, day-of-week, or seasonality; thus, clusters could be driven by these effects rather than by an out-of-control process, though adjustment for such influences could, for certain implementations, be contraindicated. Many CUSUM cluster detections reported in the relevant literature might be associated with these methodological features taken together.[4] However, caution is warranted in drawing this conclusion solely from the current analysis of this NSQIP dataset.

There was no statistically significant evidence that cluster counts increased with hospital sample size. However, it cannot be ruled out that, with ever larger sample sizes and with more opportunities for events and clusters of events, empirical and randomization cluster counts might diverge. In addition, the sensitivity analysis, using the longer 150-day window (yielding more cases in each window), did not show cluster detections at greater than chance levels, though this paradigm is an imperfect analog to 100% sampling within a 30-day window. Thus, the present findings suggest that sampling did not bias this study toward not detecting clusters at rates beyond chance, but it is not an unreasonable possibility that sensitivity to cluster detections, in excess of chance, will increase with sample size; it is a common finding that almost any "effect" can be detected at a statistically significant level with a large enough sample. However, with this NSQIP dataset, a clinically important, practical clustering "effect" was not observed.

In the absence of evidence for rapid rate shifts observed here, consideration moves to CUSUM's potential advantage in detecting gradual changes in event rates compared with ACS-NSQIP periodic reports. Two features might contribute to an advantage. First, periodic reports rely on contemporaneous modeling of 1 year of data, which requires time, while CUSUM relies on preestablished boundaries. Thus, CUSUM avoids the several months needed for fully calibrated ACS-NSQIP modeling. Second, CUSUM monitors accumulating events continuously rather than averaged over the 1-year period studied. As a result, CUSUM can detect gradual, within-hospital changes earlier than periodic reports. However, a direct comparison would require consideration of CUSUM detections possibly being false positives and reliant on the expected rate being derived from an all-hospital model (ie, the nominal early detections being driven by steady-state poor quality rather than by rate shifts of important magnitude).

CUSUM would not have the same advantages over ACS-NSQIP "on-demand" reporting that it has for traditional quarterly reports. On-demand assessments are based on historical equations and preprocessed data so that they are available almost immediately after the data collector determines that the case is complete—in a fashion probably the same as it is for CUSUM. Thus, both on-demand and CUSUM approaches would experience a similar, short time lag for the incorporation of new data. The primary difference remaining would be that on-demand addresses performance averaged over a minimum time interval of 1 month, while CUSUM tracks accumulating events. Ten events in 2 days are different from 10 events haphazardly distributed over 30 days, and CUSUM would have an advantage in detecting the former cluster—again, with some risk for false-positive flagging. However, this is the realm of detecting rapid rate changes (clusters) rather than gradual change, and the scan statistic findings suggest that these events occur at close to chance rates. It is unclear whether, over a 30-day time interval, a CUSUM or similar approach would have a meaningful early-detection advantage over ACS-NSQIP on-demand in the presence of a less dramatic shift in rate.

CUSUM has sometimes been suggested as an analytic approach that would enhance ACS-NSQIP periodic reporting.[4] While CUSUM's potential sensitivity advantages are arguable, they would come with programmatic costs. In a program such as ACS-NSQIP, there is a risk of information overload. Participants already have continuously available nonrisk-adjusted reports, quarterly risk-adjusted benchmarking reports, and risk-adjusted on-demand benchmarking reports, which provide immediate results for user-selected models. A new CUSUM or similar report, with a very different reporting structure involving cumulative observed-to-expected metrics and, depending on selected thresholds, a potential for detecting many clusters (including false positives), might contribute to signal fatigue. It would need to be clear what detection problems the added approach would solve and whether ACS-NSQIP event specifications could be enhanced (independently or in concert). If periodic and timely on-demand reports are sufficient for detecting relatively steady-state quality then, in the absence of evidence here for meaningful event clusters, additional implementation of methods such as CUSUM might not be warranted.

Of course, there is always reason to ask whether CUSUM or similar methods should be applied in other circumstances where their conduct and advantages are justified. The findings in this work are most relevant to assessments of hospital-wide performance, especially where the benefits of robust risk adjustment and standardized follow-up periods are clear. In the ACS-NSQIP framework, this would reflect broad models, including multiple case types, but would also apply to large-volume surgical specialties. However, there could be sub-groups of patients and procedures where case eligibility criteria would yield a smaller number of surgeons providing treatment, involving higher-risk procedures, using widely distributed segments of a facility's resources as well as external resources, and using less controllable resources, with multiple potential failure points. In these situations, there might be more opportunity for local system failures, whose effects could be concealed if averaged across all data within a hospital. Transplant surgery could be an example of a realm where CUSUM monitoring, which simultaneously involves the detection of rapid shifts in performance and comparison of steady-state quality (via an all-hospital-derived expected rate) might be most useful and has been successfully implemented.[8]

The decision to add CUSUM or similar assessment approaches to existing ACS-NSQIP periodic and "on-demand" benchmarking reports to enhance rapid detection of rate shifts requires careful evaluation of data structures, programmatic requirements, and implementation costs. The present findings on the apparent rarity of complication clusters inform that discussion, highlighting the importance of focusing on the true value-added information that any analytic approach contributes to the quality improvement challenge. While the focus here has been on NSQIP data, these same analytic strategies could guide quality assessment design in other areas.

REFERENCES

1. Cohen ME, Liu Y, Huffman KM, et al. On-demand reporting of risk-adjusted and smoothed rates for quality profiling in ACS NSQIP. *Ann Surg*. 2016;264:966–972.
2. Massarweh NN, Chen VW, Rosen T, et al. Comparative effectiveness of risk-adjusted cumulative sum and periodic evaluation for monitoring hospital perioperative mortality. *Med Care*. 2021;59:639–645.
3. Sun RJ, Kalbfleisch JD. A risk-adjusted O-E CUSUM with monitoring bands for monitoring medical outcomes. *Biometrics*. 2013;69:62–69.
4. Chen VW, Chidi AP, Dong Y, et al. Risk-adjusted cumulative sum for early detection of hospitals with excess perioperative mortality. *JAMA Surg*. 2023;158:1176–1183.
5. Stolwijk AM, Straatman H, Zielhuis GA. Studying seasonality by using sine and cosine functions in regression analysis. *J Epidemiol Community Health*. 1999;53:235–238.
6. Kulldorff M. SaTScan user guide for version 10.1. http://www.satscan.org/. Accessed July 30, 2024.
7. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Methods*. 1997;26:1481–1496.
8. Snyder JJ, Salkowski N, Zaun D, et al. New quality monitoring tools provided by the Scientific Registry of Transplant Recipients: CUSUM. *Am J Transplant*. 2014;14:515–523.