

Validation of Statistical Methods to Compare Cancellation Rates on the Day of Surgery

Franklin Dexter, MD, PhD*†‡, Eric Marcon, PhD§, Richard H. Epstein, MD||, and Johannes Ledolter, PhD‡

*Division of Management Consulting and Departments of †Anesthesia and ‡Health Management and Policy, University of Iowa, Iowa City; §Department of Industrial Maintenance, Jean Monnet University, Roanne Cedex, France; and ||Department of Anesthesiology, Jefferson Medical College

We investigated the validity of several statistical methods to monitor the cancellation of electively scheduled cases on the day of surgery: χ^2 test, Fisher's exact test, Rao and Scott test, Student's *t*-test, Clopper-Pearson confidence intervals, and Chen and Tipping modification of the Clopper-Pearson confidence intervals. Discrete-event computer simulation over many years was used to represent surgical suites with an unchanging cancellation rate. Because the true cancellation rate was fixed, the accuracy of the statistical methods could be determined. Cancellations caused by medical events, rare events, cases lasting longer than scheduled, and full postanesthesia or intensive care unit beds were modeled. We found that

applying Student's two-sample *t*-test to the transformation of the numbers of cases and canceled cases from each of six 4-wk periods was valid for most conditions. We recommend that clinicians and managers use this method in their quality monitoring reports. The other methods gave inaccurate results. For example, using χ^2 or Fisher's exact test, hospitals may erroneously determine that cancellation rates have increased when they really are unchanged. Conversely, if inappropriate statistical methods are used, administrators may claim success at reducing cancellation rates when, in fact, the problem remains unresolved, affecting patients and clinicians.

(Anesth Analg 2005;101:465-73)

Case cancellations on the day of surgery are generally undesirable. For hospitals in the United States of America (U.S.) not on a fixed annual budget, the lost revenue from each canceled case averages \$1430 to \$1700 USD per operating room (OR) hour plus the variable cost of performing the case (1,2). For non-U.S. hospitals and U.S. hospitals with a fixed annual budget (e.g., Veterans Affairs), canceling a case and performing it on another day increases costs to the physicians, hospital, patient, and society, even if overtime would have been required to perform the case on the day it was originally scheduled (3). For example, more than half of family members of pediatric patients miss at least 1 day at work when cases are canceled (4). Similarly, the person accompanying an adult patient often gives up a day of work. Finally, the appropriate managerial response to frequent cancellations rate on the day of surgery is to have patients arrive earlier on the day of surgery (5). This strategy allows moving up start times to avoid gaps in the OR schedule, should a preceding case be canceled. However, this strategy increases average patient

waiting times on the day of surgery (5), which may decrease patient satisfaction.

There have been many research studies evaluating causes of cancellations on the day of surgery (e.g., 6-10). However, actually monitoring case cancellation rates and determining change over time or differences among specialties is difficult. For example, when the intensive care unit (ICU) fills, there are many cancellations both for services whose patients require postoperative ICU care and for services using the postanesthesia care unit (PACU), the ICU overflow site. If the ICU fills once or twice a month, and cancellations are being compared from one month to the next, the cancellation rate may seem to vary markedly from month to month, leading to poor management decisions.

In research studies, when statistical methods are used to compare cancellations among groups, often the χ^2 test is chosen (6-8). Confidence intervals for odds ratios are estimated by logistic regression to adjust for patients' baseline characteristics (9,10). These methods are fine for analyses of patient risk of medical events because each patient's risk of a medical event is statistically independent of all other patients' risks of the event. When the risk of one patient's case being canceled is correlated to that of another patient, such methods break down.

Accepted for publication December 13, 2004.

Address correspondence and reprint requests to Franklin Dexter, Anesthesia 6-JCP, University of Iowa, Iowa City, Iowa 52242. Address e-mail to franklin-dexter@uiowa.edu.

DOI: 10.1213/01.ANE.0000154536.34258.A8

Table 1. Simulated Types of Cancellations of Surgical Cases

Type	Description
Medical	Fact that one patient has surgery canceled does not affect another patient's risk of their surgery being canceled. For example, a patient develops acute sinusitis within a week of elective surgery. For example, a patient develops a pulmonary embolism on the morning planned for surgery.
Rare Event	Single events that can cause more than one case to be canceled. For example, a new, part-time nurse in the otolaryngology clinic gives patients incorrect fasting information. For example, a surgeon is sick. For example, PACU fills infrequently, but when it happens, delays are so long that several cases are canceled. For example, an unusually large number of urgent cases occur during scheduled hours.
Cases running late	This cause would include the circumstances of a preceding case in an OR taking much longer than expected. This cause would also include a surgeon who follows another surgeon in the same OR on the same day, but is delayed in arriving from another site.
Full PACU	Delays in discharge from ORs because of a full PACU, intensive care unit, phase II PACU, and/or hospital ward resulting in cases being canceled in multiple ORs. This type of cancellation represents any cause for which there is a correlation in risks among services.

OR = operating room; PACU = postanesthesia care unit.

In the reality of clinicians' and managers' surgical suites, many cancellations result from nonmedical causes (e.g., full ICU, full PACU, surgeon unavailable, bad weather, or urgent cases). Whenever one of these nonmedical causes occurs, more than one case can be canceled. For example, at a university hospital with outpatient preoperative evaluation, when adults had their surgery canceled on the day of surgery, nonmedical causes were responsible for 80% of cancellations (9). At a Veterans Affairs Hospital, nonmedical reasons accounted for 67% of cancellations before the introduction of outpatient preoperative evaluations (6) and 81% of cancellations after a year of experience with this process (7). Among inpatients, 43% of cancellations were caused by nonmedical factors (11). Among all patients at a tertiary teaching hospital, 68% of cancellations had nonmedical causes (12). The issue likely is less relevant to pediatrics because percentages for nonmedical causes of case cancellation are lower than for adults: 15% in one study (10) and 33% in another (4).

We studied several statistical methods for analyzing case cancellations to determine which methods can be used accurately for clinicians and managers' routine monitoring needs. In the Discussion, we include a worked example demonstrating the recommended method so that readers can easily implement the findings of the study.

Methods

Type I and Type II Error Rates

Type I errors occur when there are no true differences between groups, and yet statistically significant differences are detected. If the nominal chance of a type I

error (α) is set equal to 0.05 (i.e., $P < 0.05$ is significant), a test should not achieve significance more often than on 5% of occasions unless there are true differences between groups. Because decisions based on faulty analysis often result in the implementation of processes that waste everyone's time (e.g., additional paperwork, phone calls, and laboratory and diagnostic testing), these type I errors can have a detrimental effect. Similarly, a type I error may lead some administrators to claim success at reducing cancellation rates when, in fact, there has been no change.

Type II errors occur when significant differences are not detected, even though there are true differences between groups. Statistical power is high when type II error rates are low. For example, type II errors occur when some services suffer from full ICUs, but statistical tests show those cancellations do not differ significantly from those of other services. Evaluation of type II errors is relevant provided statistical methods have appropriate type I error rates.

Descriptions of Statistical Methods

Cancellations caused by medical events (Table 1) were used to represent all types of cancellations for which the fact that one patient was canceled does not change the probability that other patients have their surgeries canceled. Mr. Jones developing chest pain in the holding area before his inguinal hernia repair does not influence the probability that Mrs. Smith will have an increased temperature and white count before her total hip replacement.

Fisher's exact test will have an appropriate type I error rate (i.e., equal to its nominal, correct value)

Table 2. Frequencies of Occurrence of Types of Cancellations Listed in Table 1

Suite	Type	Cases affected (%)	Days with at least one such cancellation (%)
5 OR	Medical	0.80	16
	Rare event	0.99	7.9
	Cases running late	3.06	49
	Full PACU	0.83	16
	Total	5.68	69
15 OR	Medical	0.80	37
	Rare event	0.99	23
	Cases running late	3.71	89
	Full PACU	0.90	41
	Total	6.40	98

The table shows results of 2 simulations, 1 for 5 operating rooms (ORs) and the other 15 ORs. The overall cancellation rates are reported for the 5 OR and 15 OR surgical suites, without regard to service. Each simulation included 65,000 4-wk periods (i.e., 5,200 simulated yrs). All standard errors are less than 0.02%. The values in the first column sum to the Total because each case was canceled for just one reason. The values in the second column do not sum to the Total because each day could have several cancellations of different types.

PACU = postanesthesia care unit.

when comparing rates of cancellations from medical events between groups (e.g., between services or between 6-mo periods). If $P = 0.05$ is considered significant, then 5% of comparisons should demonstrate a statistical change purely based on chance. The χ^2 test will behave similarly, provided there are at least five cancellations in each of the groups being compared. The χ^2 test can be performed in a spreadsheet such as Excel using built-in functions. A corresponding method for calculating confidence intervals for proportions is the method of Clopper-Pearson (13), implementable in Excel as one formula.

Statistical methods to analyze nonmedical causes of cancellations can consider variations in cancellation rates within and among short periods (14). The principal determinant of OR workload by subspecialty is the day of the week (15,16). Vacations, meetings, variations in clinics, etc., are often 2 weeks long. Consequently, we considered 4 weeks the shortest data collection period that would be used without considering variation by day of the week (17-20). Our choice of 4 weeks was similar to previously published periods of multiples of months: 1 mo (8), 3 mo (7,9,10), 4 mo (12), and 6 mo (6,11).

The statistical methods estimate the variance in cancellation rates among different 4-week periods and add it to the estimate of the variance in cancellation rates among cases within the same period. The Rao and Scott method (21) has the highest statistical power among competing methods for comparing two groups, without exceeding nominal rates (21-23). Chen and Tipping (24) described an analogous method for modifying Clopper-Pearson confidence intervals. We used sets of six 4-week periods for our comparisons. A sample size of six is small statistically, but even that duration is the longest period pooled in practice when studying cancellations (6-12).

Alternatively, the uncertainty in the true percentage cancellation rate within each of the 4-week periods can

be ignored (25), and Student's two-sample t -test with unequal variances applied to 2 samples of 6 numbers each. Confidence intervals for the means of single sets of six 4-week periods are calculated with the Student t distribution (Appendix). This approach has been used widely for the statistical analysis of other OR management data, including staffing costs (17,18), ORs in use at different times of the day (19), and OR workload for purposes of OR allocation (20). However, those values are not percentage cancellation rates with values that can be close to zero. The method may work poorly when percentages are nearly equal to zero. Consequently, we followed Shirley and Hickling (25) in using the Student's t -test after transforming the percentages (26), using Equation (1) of the Appendix.

Testing Statistical Methods

The validity of statistical methods is evaluated using computer simulation. A set of real data can be used to investigate whether different statistical methods give the same answer, but that does not show which answer, if either, is correct. The underlying statistical distribution used to generate the real data would be unknown, and the real data would be only one realization of the underlying statistical distribution.

We used the above referenced research (3,4,6-9,12) and other papers to assure that the conditions simulated were realistic (Appendix). Computer simulation provided known, correct answers to which the results of the statistical methods could be compared.

Simulated data to test the statistical methods were obtained using ARENA version 7.01 (Rockwell Software, Sewickley, PA). For each of eight different combinations of parameter values (Appendix; Tables 2-7), simulation output was counts of canceled and noncanceled cases for 65,000 4-week periods of 20 workdays. Because the cancellation rate was fixed over these

Table 3. Significant Change in Cancellation Rate Comparing 1 4-wk Period to the Next Based on a Nominal (Correct) Type I Error Rate of 5%

Suite	Types of cancellations simulated	χ^2	Fisher's exact test
5 OR	All four types	7.7% ± 0.1%	6.7% ± 0.1%
	Medical events only	4.6% ± 0.1%	2.3% ± 0.1%
	Rare events only	30.9% ± 0.3%	22.3% ± 0.2%
	Rare events excluded	3.6% ± 0.1%	2.8% ± 0.1%
15 OR	All four types	6.4% ± 0.1%	5.9% ± 0.1%
	Medical events only	4.9% ± 0.1%	3.7% ± 0.1%
	Rare events only	26.0% ± 0.2%	23.5% ± 0.2%
	Rare events excluded	3.6% ± 0.1%	3.3% ± 0.1%

If a test performed perfectly, 5% of comparisons would be detected as significantly different. Values larger than 7% are marked in bold. Standard errors (±) were estimated by applying Clopper-Pearson method to each simulated period of 40 workdays. Each of the 8 separate simulation studies included 65,000 4-wk periods (i.e., 5,200 simulated yr). The overall cancellation rates were tested statistically for the 5 operating room (OR) and 15 OR surgical suites, without regard to service.

Table 4. Significant Change in Cancellation Rate Comparing Data from 6 4-wk Periods to Data from the Next 6 4-wk Periods Based on a Nominal (Correct) Type I Error Rate of 5%

Suite	Types of cancellations simulated	χ^2	Rao and Scott	Student's two-sample <i>t</i> -test	Student's two-sample <i>t</i> -test with Freeman-Tukey transformation
5 OR	All four types	7.4% ± 0.4%	4.9% ± 0.3%	4.5% ± 0.3%	4.5% ± 0.3%
	Medical events only	4.8% ± 0.3%	3.5% ± 0.3%	4.9% ± 0.3%	4.8% ± 0.3%
	Rare events only	27.8% ± 0.6%	8.1% ± 0.4%	3.9% ± 0.3%	4.5% ± 0.3%
	Rare events excluded	3.2% ± 0.2%	2.5% ± 0.2%	4.6% ± 0.3%	4.5% ± 0.3%
15 OR	All four types	5.4% ± 0.2%	3.6% ± 0.3%	4.2% ± 0.3%	4.1% ± 0.3%
	Medical events only	4.3% ± 0.3%	2.9% ± 0.2%	4.0% ± 0.3%	4.2% ± 0.3%
	Rare events only	25.6% ± 0.6%	7.6% ± 0.4%	4.5% ± 0.3%	4.8% ± 0.3%
	Rare events excluded	3.7% ± 0.3%	3.0% ± 0.2%	4.9% ± 0.3%	4.8% ± 0.3%

If a test performed perfectly, 5% of comparisons would be detected as significantly different. Values larger than 7% are marked in bold. Standard errors (±) were estimated by applying Clopper-Pearson method to each simulated period of 240 workdays. Each of the 8 simulations included 65,000 4-wk periods (i.e., 5,200 simulated yr). The overall cancellation rates were tested statistically for the 5 operating room (OR) and 15 OR surgical suites, without regard to service.

Table 5. Significantly Different Cancellation Rate of One Service Versus Others Using 1 and 6 4-wk Periods of Data and Nominal (Correct) Type I Error Rate of 5%

Suite	Types of cancellations simulated	χ^2 one 4-wk period	Fisher's exact test one 4-wk period	χ^2 six 4-wk periods	Rao and Scott	Student's two-sample <i>t</i> -test	Student's <i>t</i> Freeman Tukey
5 OR	All four types	5.8% ± 0.1%	4.4% ± 0.1%	6.3% ± 0.2%	4.1% ± 0.2%	4.6% ± 0.2%	4.7% ± 0.2%
	Medical events only	4.4% ± 0.1%	1.3% ± 0.1%	4.7% ± 0.2%	3.6% ± 0.2%	4.9% ± 0.2%	5.5% ± 0.2%
	Rare events only	23.9% ± 0.2%	13.2% ± 0.1%	26.6% ± 0.4%	8.1% ± 0.3%	4.2% ± 0.2%	4.5% ± 0.2%
	Rare events excluded	3.4% ± 0.1%	2.3% ± 0.1%	3.3% ± 0.2%	2.6% ± 0.2%	4.5% ± 0.2%	4.5% ± 0.2%
15 OR	All four types	4.9% ± 0.1%	4.3% ± 0.1%	5.1% ± 0.2%	3.5% ± 0.2%	4.6% ± 0.2%	4.8% ± 0.2%
	Medical events only	4.0% ± 0.1%	2.8% ± 0.1%	4.9% ± 0.2%	3.5% ± 0.2%	5.0% ± 0.2%	5.4% ± 0.2%
	Rare events only	20.4% ± 0.2%	16.9% ± 0.2%	21.3% ± 0.4%	7.6% ± 0.3%	4.8% ± 0.2%	5.1% ± 0.2%
	Rare events excluded	3.4% ± 0.1%	2.8% ± 0.1%	3.2% ± 0.2%	2.4% ± 0.2%	4.7% ± 0.2%	4.6% ± 0.2%

If a test performed perfectly, 5% of comparisons would be detected as significantly different. Values larger than 7% are marked in bold. The Table was created by switching the mean case duration of the 1-h and 3-h services to 2-h, such that there was no true difference between the 2-h service and the other 2 services (see Appendix). Standard errors (±) were estimated by applying Clopper-Pearson method to each simulated period of set of 4-wk periods. Each of the 8 simulations included 65,000 4-wk periods (i.e., 5,200 simulated yr).

OR = operating rooms.

5,200 years of data, we could evaluate whether statistical tests would have a type I error rate exceeding 5% (the expected value) with a $P < 0.05$ criterion.

Visual Basic for Excel 2003 (Microsoft, Redmond, WA) was used for statistical analysis of the output of the ARENA simulations. Cancellations because of

Table 6. Percentages of 95% Confidence Intervals that Failed to Include the True Cancellation Rates with 6 4-wk Periods of Data

Suite	Types of cancellations simulated	Clopper-Pearson	Clopper-Pearson corrected for variance inflation	Student <i>t</i> one sample	Student <i>t</i> one sample with Freeman-Tukey transformation
5 OR	All four types	7.0% ± 0.2%	4.6% ± 0.2%	4.9% ± 0.2%	4.9% ± 0.2%
	Medical events only	3.6% ± 0.2%	2.8% ± 0.2%	5.9% ± 0.2%	6.2% ± 0.2%
	Rare events only	24.8% ± 0.4%	8.3% ± 0.3%	7.3% ± 0.3%	5.4% ± 0.2%
	Rare events excluded	2.9% ± 0.2%	2.4% ± 0.1%	4.7% ± 0.2%	5.2% ± 0.2%
15 OR	All four types	5.8% ± 0.2%	4.1% ± 0.2%	4.9% ± 0.2%	4.9% ± 0.2%
	Medical events only	4.2% ± 0.2%	3.0% ± 0.2%	5.1% ± 0.2%	5.2% ± 0.2%
	Rare events only	24.6% ± 0.4%	8.7% ± 0.3%	5.7% ± 0.2%	4.9% ± 0.2%
	Rare events excluded	3.1% ± 0.2%	2.4% ± 0.2%	5.2% ± 0.2%	5.3% ± 0.2%

If a test performed perfectly, 5% of the 95% confidence intervals would not include the true cancellation rate. Values larger than 7% are marked in bold. Standard errors (±) were estimated by applying Clopper-Pearson method to each simulated period of 120 workdays. Each of the 8 simulations included 65,000 4-wk periods (i.e., 5,200 simulated yr). The overall cancellation rates were tested statistically for the 5 operating room (OR) and 15 OR surgical suites, without regard to service.
OR = operating room.

Table 7. Significantly (*P* < 0.05) Different Cancellation Rate of One Service Versus Others Using 6 4-wk Periods of Data When Services Truly Differ: Statistical Power Analysis

Suite	Types of cancellations simulated	Student's two sample <i>t</i> -test	Student's two sample <i>t</i> -test with Freeman-Tukey transformation
5 OR	All four types	13.6% ± 0.3%	17.9% ± 0.4%
	Rare events excluded	20.0% ± 0.4%	26.9% ± 0.4%
15 OR	All four types	26.1% ± 0.4%	30.5% ± 0.4%
	Rare events excluded	35.8% ± 0.5%	41.2% ± 0.5%

Because services differed in their mean scheduled durations, the number of cases per operating room (OR) per day and the number of cases per surgeon per day differed among services. Consequently, the proportion of cases canceled for nonmedical reasons differed among services. The cancellation rate for the service with mean scheduled duration of 2.0 h was compared to the cancellation rate for the other 2 services combined. The null hypothesis of no difference in cancellation rates was tested against the two-sided alternative using 120 workdays in each of 2 groups. Standard errors (±) were estimated by applying Clopper-Pearson method to each simulated period of 120 workdays. Each of the 4 simulations included 65,000 4-wk periods (i.e., 5,200 simulated yr).

medical events alone were used to confirm our computer code because we knew all methods would perform well for these simulations. The Discussion includes limitations of the simulations and an example using real OR data.

Results

We simulated 5 OR and 15 OR surgical suites to represent small and large facilities, respectively. Cancellation rates were 5.7% for the 5 OR surgical suites and 6.4% for the 15 OR surgical suites (Table 2). Rare events caused the cancellation of 1.0% of scheduled cases by causing events on 7.9% of days at the 5 OR surgical suites and 23% of days at the 15 OR surgical suites.

When testing for differences from one 4-week period to the next, both the χ^2 test and Fisher's exact test had high type I error rates caused by rare events (Table 3). Results were similar when comparing six 4-week periods to the next six 4-week periods (Table 4) and when comparing cancellation rates between services (Table 5). The type I error rate for all four types of errors combined represented the mixture between achieving statistical significance too often when

rare events were present and too infrequently from other causes (Tables 3–5). Likewise, Clopper-Pearson confidence intervals included the true rate of cancellations caused by rare events far too infrequently (Table 6).

Rao and Scott (21) and Chen and Tipping (24) methods were more accurate than the χ^2 and Clopper-Pearson methods, respectively, but still had type I error rates exceeding the nominal value of 5% when cancellations were caused by rare events (Tables 4 and 6). When all four types of cancellations were present, performance was sensitive to the incidence of cancellations caused by rare events.

Student's *t*-test and analogous methods were generally accurate (Tables 4–6). The same finding was obtained when the counts were first transformed. The latter method had the smallest absolute difference from the expected 5% type I error rate for confidence intervals in the simulation of the five OR surgical suites with cancellations caused by rare events only (Table 6). In that circumstance, 19% of the 4-week periods had no observed cancellations, and 60% had 4 or less (see Discussion).

Table 7 studies type II errors, as described in the first section of Methods. Statistical power to detect

Table 8. Example of Applying Student’s *t* Test with Freeman-Tukey Double Arcsine Transformation for Comparing Cancellation Rates between Two Services

Canceled cases during the first 120 workdays of 2003 at the tertiary surgical suite of a United States academic hospital						
Four-wk period:	1st period Jan 2 to Jan 30	2nd period Jan 31 to Feb 27	3rd period Feb 28 to Mar 27	4th period Mar 28 to Apr 24	5th period Apr 25 to May 22	6th period May 23 to Jun 20
General surgery						
Canceled	45	32	50	52	51	38
Scheduled	378	359	349	367	365	392
All other services						
Canceled	80	66	66	76	71	69
Scheduled	837	788	800	859	813	831
Percentages of cases canceled before and after applying Freeman-Tukey double arcsine transformation (equation 1)						
General surgery						
Cancellation rate	11.9%	8.9%	14.3%	14.2%	14.0%	9.7%
Transformed value	0.35	0.31	0.39	0.39	0.38	0.32
All other services						
Cancellation rate	9.6%	8.4%	8.3%	8.8%	8.7%	8.3%
Transformed value	0.32	0.29	0.29	0.30	0.30	0.29
Calculation of sample statistics from the transformed values						
General surgery	<i>n</i> = 6, sample mean = 0.356, sample SD = 0.037					
All other services	<i>n</i> = 6, sample mean = 0.300, sample SD = 0.009					
Statistical analysis using Student’s <i>t</i> -test with unequal standard deviations						
$t = \frac{\sqrt{6} (0.356 - 0.300)}{\sqrt{0.037^2 + 0.009^2}}$						
$\text{degrees of freedom} = \frac{(6 - 1) (0.037^2 + 0.009^2)^2}{0.037^4 + 0.009^4}$						
<i>P</i> -value = 0.013						

differences in cancellation rates between services was significantly higher for Student’s *t*-test applied to transformed counts than without transformation.

Discussion

Statistical methods used in research of medical causes of events (e.g., Fisher’s exact test) should not be used by clinicians and managers in their routine monitoring of case cancellations because of their high type I error rates. Internal and benchmarking quality reports should use Student’s *t*-test and analogous methods applied to cancellation rates from 4-week periods after transforming the counts.

Table 8 provides an example of the method using real data from an academic medical center. Table 8 also shows the usefulness of the method. The method can be implemented in a few lines of computer code and a spreadsheet (e.g., Excel). A manager can test the answer provided to him or her by computer software using small amounts of data (e.g., that in Table 8). Finally, the method is based simply on the numbers of

canceled versus performed cases. Although we studied effects of different types of cancellations in this paper, the usefulness of the method is unaffected by the ability of a facility to track and categorize the reason for each of its case cancellations.

Different Statistical Methods

We do not recommend excluding days with cancellations caused by rare events for three reasons. First, administrators can be motivated to reduce cancellations because of their economic importance. Excluding days with rare events shows smaller benefit to preventing cancellations. For example, if anesthesiologists want to show that transplant cases occurring early on weekdays markedly disrupt the elective schedule, excluding those days from the report makes no sense. Second, our experience is that many rare events are not caused by snowstorms but rather events that are hard to identify clearly in practice. For example, although a surgeon’s flight home may be delayed, resulting in the cancellation of his cases scheduled for the next day, that reason may not be

reported in the OR log sheet. Third, the definitions and types of rare events are likely to vary among facilities and or surgical populations. Trying to set systematic and valid policies for exclusion of rare events will be challenging. The consequence of not excluding rare events is that precisely what types of cancellations are rare do not need to be defined.

We do not recommend using Fisher's exact test or similar methods to compare cancellation rates when review of the data suggests that few of the observed cancellations were caused by rare events. A 1% cancellation rate attributable to rare events (Table 2) was sufficient to affect statistical methods markedly (Tables 3-6). Some surgical suites will have an incidence of cancellations caused by rare events of <1%. Yet, they are unlikely to know their true incidence because the upper bound on the incidence of cancellations from rare events cannot be estimated accurately using methods appropriate for medical events (Table 6). Thus, we recommend simply using Student's *t*-test applied to transformed data for OR cancellations.

Rao and Scott and Chen and Tipping methods performed worse than we expected (21-24). Our results probably differed from those previously reported because the previous papers used sample sizes applicable to toxicology studies, not case cancellations. First, we studied only six 4-week periods versus toxicology studies with 30 or so litters of pups, for which those methods perform well. Our sample size of six was probably too small for accurate estimation of the variances in cancellation rates among 4-week periods. Second, we had hundreds of scheduled cases within each 4-week period versus toxicology with litters of 2-12 pups. Consequently, there was relatively little uncertainty in cancellation rates within 4-week periods, just uncertainty among periods. This pattern explains why Student's *t*-test and analogous methods performed quite well.

We did not study nonparametric methods such as Mann-Whitney-Wilcoxon (23) because parametric methods like Student's *t*-test have higher statistical power, and, for our application, performed well after data transformation (Tables 4 and 5).

Limitations

Our results are likely valid because they seem unaffected by the characteristic of our mathematical model of a surgical suite other than with respect to the incidence of a rare event and the resulting number of case cancellations from each rare event. Thus, our results apply fully to the many surgical suites with virtually no daily cancellations because of cases running late or full PACUs. Fine tuning our model to be more realistic for any one surgical suite would likely be of little or no value other than to the extent that we more realistically model the characteristics of rare cancellations at

the specific suite. The pattern of such events likely varies depending on each suite's unique circumstances, such that additional realism would make results less valid for most other sites. This limitation is moot provided a statistical method is used that is robust to rare events. That is why we recommend that (almost) all facilities that monitor cancellation rates use such a method (e.g., as in Table 8).

Cancellation rates likely vary among facilities, depending partly on the types of patients receiving care. For example, some published cancellation rates (including those on the day before surgery) include 4.6% for outpatients (9), 6.6% for outpatients (6), 9% for outpatients (11), 10% among outpatients (12), 10% among pediatric outpatients (10), 12% among plastic surgery patients (8), 13% overall (7), 17% among inpatients (11), 19% among inpatients (6), and 30% among inpatients (12). We studied cancellation rates on the day of surgery between 0.8% and 6.4% (Table 2). We recommend that our results not be applied by facilities lacking at least one observed cancellation in each of the 6 studied 4-week periods. We repeated the simulations with just rare events, using only two ORs and only the service with two-hour average case durations. Confidence intervals were created using Student's *t* distribution with the transformation applied to six 4-week periods, as in Table 6. The 95% confidence intervals failed to contain the true cancellation rate for $11.9\% \pm 0.3\%$ of comparisons. This unacceptably high type I error rate occurred because 57% of 4-week periods had no cancellations. We doubt that our inability to consider less than one cancellation every four weeks is a major limitation, because when the incidence is so infrequent, most clinicians and managers would be uninterested in quantifying cancellations.

Summary

Clinicians and managers interested in routine monitoring of OR cancellation rates generally need a robust method that can be applied automatically, without a formal statistical assessment like a research study. We recommend calculating the number of canceled and performed cases during each four-week period, transforming each period's cancellation rate, and then applying Student's *t*-test. Methods such as Fisher's exact test and χ^2 test can give highly misleading results, resulting in inappropriate management decisions.

Appendix

Computer Simulation

Discrete-event computer simulation (27) was used to represent the random flow of patients from ORs

through the PACU. Each workday was simulated independently of all other workdays. Simulation was performed for 5 OR and 15 OR surgical suites.

Scheduled case durations were described using different log-normal distributions for each of three services. Each service had a mean scheduled duration of 1.0, 2.0, or 3.0 h, with a common standard deviation of the logarithm of case duration in hours equal to 0.725 (28). After calculation, the scheduled durations were bounded between 0.3 and 1.9 h for the 1-h service, between 0.6 and 3.9 h for the 2-h service, and between 0.9 and 5.9 h for the 3-h service. The actual case durations were calculated using the method described by Kennedy (29) to include the differences between scheduled and actual case durations that were measured by Goldman et al (30). Specifically, actual case durations were set equal to the scheduled case duration multiplied by a normally distributed random number with a mean of 1.00 and sd of 0.25 (31).

Each turnover time ("patient out" to "patient in") was assigned a time duration generated randomly from a log-normal distribution with mean \pm sd = 0.30 \pm 0.20 h, bounded between 0.17 and 1.50 h.

Each OR in the surgical suite had two surgeons. The first surgeon completed his or her cases, followed by the second surgeon. The cases were divided randomly, with equal probability, between the two surgeons. Often this resulted in an unequal number of cases performed by the two surgeons in each OR. For the 5 OR and 15 OR surgical suites, 2 ORs and 5 ORs were allocated for the service with a mean duration of 1.0 h, respectively. Cases were scheduled sequentially using an 8-h workday. Adjusted use (OR time plus turnovers) was 83.7% \pm 0.1% (SE). For the 5 OR and 15 OR surgical suites, 2 ORs and 5 ORs were allocated for the service with a mean duration of 2.0 h. Adjusted use was 77.6% \pm 0.1%. For the 5 OR and 15 OR surgical suites, 1 OR and 5 ORs were allocated for the service with a mean duration of 3.0 h. Adjusted use was 71.2% \pm 0.1%.

Cancellations caused by rare events were represented by the unexpected absence of a surgeon. Whether a surgeon was unavailable was determined by a Bernoulli distributed random number. If the surgeon was unavailable, all of that surgeon's cases for the day, from the preceding paragraph, were canceled. The achieved risk of any one case being canceled from this cause was 1.0% (Table 2).

Cancellations caused by medical causes were simulated by generating a Bernoulli distributed random number with a 0.8% probability. Such cancellations occurred equally frequently for the three services, unlike cancellations caused by other causes.

Cancellations caused by cases running late were used to represent cancellations from any cause providing correlation in risks within services. If a case was expected, from its scheduled duration, to finish

more than 0.5 h after the end of the 8-h workday, the case was canceled.

Cancellations caused by a full PACU were used to represent cancellations from any cause providing correlation in risks among services. Ten PACU beds were planned for the 5 OR surgical suite and 30 PACU beds for the 15 OR suite. Each patient's time in the PACU was generated from a lognormal statistical distribution with a mean of 1.0 h and sd of 1.2 h, bounded between 0.5 and 3.0 h. If the PACU was full, discharges from ORs into the PACU were delayed in original sequence. A case was canceled if the patient was expected to enter the PACU more than 1.5 h after the end of the 8-h workday.

Whether a case was canceled was determined in the sequence of medical cause, rare event, cases running late, and then full PACU.

Freeman-Tukey Double Arcsin Transformation

The Freeman-Tukey double arcsin transformation (26) equals

$$\theta = \frac{1}{2} \left(\arcsin \left[\sqrt{\frac{c}{n+1}} \right] + \arcsin \left[\sqrt{\frac{c+1}{n+1}} \right] \right), \quad (1)$$

where c is the number of cancellations and n is the number of scheduled cases during a 4-week period. Table 8 gives an example of applying the transformation.

The inverse of the transformation is required only when calculating 95% confidence intervals for the cancellation rate, as in Table 6. Calculate the sample mean $\bar{\theta}$ and sd s_{θ} of the transformed values $\theta_1, \theta_2, \dots, \theta_p$ from each of the p four-week periods. Estimate confidence intervals by

$$\bar{\theta} \pm \frac{s_{\theta} t_{1-0.05/2, p-1}}{\sqrt{p}}, \quad (2)$$

where t is the inverse of the Student t-distribution with p-1 degrees of freedom. Report the value of $\bar{\theta}$ and its confidence intervals after taking the inverse of the transformation of Equation (1).

To calculate the inverse of the transformation, we use the bisection method in our Visual Basic for Excel code (32). For convenience, we show the steps for $\bar{\theta}$. The same steps are applied to the lower and upper intervals.

1. $U = N = n_1 + n_2 + \dots + n_p$
2. $L = 0$
3. $c = L$
4. $s = U - L$
5. $s = s/2$
6. $U = c + s$

$$7. Z = -\bar{\theta} + \frac{1}{2} \left(\arcsin \left[\sqrt{\frac{U}{N+1}} \right] + \arcsin \left[\sqrt{\frac{U+1}{N+1}} \right] \right)$$

8. Select Case

- $Z = 0$ or $s < 0.000001$
Return c/N as the actual cancellation rate
- $Z < 0$
 $c = U$
Repeat steps 5 to 8
- $Z > 0$
Repeat steps 5 to 8

Readers can check their implementation of the steps by using the transformed and nontransformed cancellation rates in Table 8.

References

1. Macario A, Dexter F, Traub RD. Hospital profitability per hour of operating room time can vary among surgeons. *Anesth Analg* 2001;93:669-75.
2. Dexter F, Blake JT, Penning DH, Lubarsky DA. Calculating a potential increase in hospital margin for elective surgery by changing operating room time allocations or increasing nursing staffing to permit completion of more cases: a case study. *Anesth Analg* 2002;94:138-42.
3. Tessler MJ, Mitmaker L, Wahba RM, Covert CR. Patient flow in the postanesthesia care unit: an observational study. *Can J Anaesth* 1999;46:348-51.
4. Tait AR, Voepel-Lewis T, Munro HM, et al. Cancellation of pediatric outpatient surgery: economic and emotional implications for patients and their families. *J Clin Anesth* 1997;9:213-9.
5. Dexter F, Traub RD. Statistical method for predicting when patients should be ready on the day of surgery. *Anesthesiology* 2000;93:1107-14.
6. Pollard JB, Zboray AL, Mazze RI. Economic benefits attributed to opening a preoperative evaluation clinic for outpatients. *Anesth Analg* 1996;83:407-10.
7. Pollard JB, Olson L. Early outpatient preoperative anesthesia assessment: does it help to reduce operating room cancellations? *Anesth Analg* 1999;89:502-5.
8. Guyuron B, Zarandy S. Causes for cancellation of aesthetic and reconstructive procedures. *Plast Reconstr Surg* 1993;92:662-70.
9. van Klei WA, Moons KGM, Rutten CLG, et al. The effect of outpatient preoperative evaluation of hospital inpatients on cancellation of surgery and length of hospital stay. *Anesth Analg* 2002;94:644-9.
10. Macarthur AJ, Macarthur C, Bevan JC. Determinants of pediatric day surgery cancellation. *J Clin Epidemiol* 1995;48:485-9.
11. Hand R, Levin P, Stanziola A. The causes of cancelled elective surgery. *Qual Assur Util Rev* 1990;5:2-6.
12. Lacqua MJ, Evans JT. Cancelled elective surgery: an evaluation. *Am Surg* 1994;60:809-11.
13. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;17:857-72.
14. Collett D. *Modelling binary data*. London: Chapman & Hall, 1991;188-194.
15. Strum DP, Vargas LG, May JH. Surgical subspecialty block utilization and capacity planning: a minimal cost analysis model. *Anesthesiology* 1999;90:1176-85.
16. Dexter F, Epstein RH, Marsh HM. Statistical analysis of week-day operating room anesthesia group staffing at nine independently managed surgical suites. *Anesth Analg* 2001;92:1493-8.
17. Abouleish AE, Dexter F, Epstein RH, et al. Labor costs incurred by anesthesiology groups because of operating rooms not being allocated and cases not being scheduled to maximize operating room efficiency. *Anesth Analg* 2003;96:1109-13.
18. Dexter F, Abouleish AE, Epstein RH, et al. Use of operating room information system data to predict the impact of reducing turnover times on staffing costs. *Anesth Analg* 2003;97:1119-26.
19. Dexter F, Epstein RH. Optimizing second shift OR staffing. *AORN J* 2003;77:825-30.
20. Dexter F, Macario A, Qian F, Traub RD. Forecasting surgical groups' total hours of elective cases for allocation of block time: application of time series analysis to operating room management. *Anesthesiology* 1999;91:1501-8.
21. Rao JNK, Scott AJ. A simple method for the analysis of clustered binomial data. *Biometrics* 1992;48:577-85.
22. Paul SR, Islam AS. Analysis of proportions in the presence of over- under-dispersion. *Biometrics* 1995;51:1400-10.
23. Lee S. Analysis of the binary littermate data in the one-way layout. *Biomtrc J* 2003;45:195-206.
24. Chen C, Tipping RW. Confidence interval of a proportion with over-dispersion. *Biomtrc J* 2002;44:877-86.
25. Shirley EAC, Hickling R. An evaluation of some statistical methods for analyzing numbers of abnormalities found amongst litters in teratology studies. *Biometrics* 1981;37:819-29.
26. Mosteller F, Youtz C. *Tables of the Freeman-Tukey transformations for the binomial and Poisson distributions*. *Biometrika* 1961;48:433-40.
27. Banks J, Carson J, Nelson B. *Discrete-event system simulation*. 3rd ed. Paramus, NJ: Prentice Hall, 2000:1-40.
28. Dexter F, Macario A, Manberg PJ, Lubarsky DA. Computer simulation to determine how rapid anesthetic recovery protocols to decrease the time for emergence or increase the phase I post anesthesia care unit bypass rate affect staffing of an ambulatory surgery center. *Anesth Analg* 1999;88:1053-63.
29. Kennedy, MH. *Bin-packing, knapsack, and change-constrained approaches to operating room scheduling [dissertation]*. Troy, New York: Rensselaer Polytechnic Institute, Department of Decision Sciences and Engineering Systems, 1992:83-86, 90, 99.
30. Goldman J, Knappenberger HA, Shearon WT. A study of variability of surgical estimates. *Hosp Manage* 1970;110:46-D.
31. Dexter F, Macario A, O'Neill L. Scheduling surgical cases into overflow block time: computer simulation of the effects of scheduling strategies on operating room labor costs. *Anesth Analg* 2000;90:980-6.
32. Press WH, Teukolsky SA, Vetterling WT, et al. *Numerical recipes in FORTRAN: the art of scientific computing*. 2nd ed. Cambridge, MA: Cambridge University Press, 1992:340-7.