

# Influence of Procedure Classification on Process Variability and Parameter Uncertainty of Surgical Case Durations

Franklin Dexter, MD, PhD,\* Elisabeth U. Dexter, MD, FACS,† and Johannes Ledolter, PhD‡

**BACKGROUND:** Predictive variability of operating room (OR) times influences decision making on the day of surgery including when to start add-on cases, whether to move a case from one OR to another, and where to assign relief staff. One contributor to predictive variability is process variability, which arises among cases of the same procedure(s). Another contributor is parameter uncertainty, which is caused by small sample sizes of historical data.

**METHODS:** Process variability was quantified using absolute percentage errors of surgeons' bias-corrected estimates of OR time. The influence of procedure classification on process variability was studied using a dataset of 61,353 cases, each with 1 to 5 scheduled and actual Current Procedural Terminology (CPT) codes (i.e., a standardized vocabulary). Parameter uncertainty's sensitivity to sample size was quantified by studying ratios of 90% prediction bounds to medians. That studied dataset of 65,661 cases was used previously to validate a Bayesian method to calculate 90% prediction bounds using combinations of surgeons' scheduled estimates and historical OR times.

**RESULTS:** (1) Process variability differed significantly among 11 groups of surgical specialty and case urgency ( $P < 0.0001$ ). For example, absolute percentage errors exceeded the overall median of 22% for 57% of urgent spine surgery cases versus 42% of elective spine surgery cases. (2) Process variability was *not* increased when scheduled and actual CPTs differed ( $P = 0.23$  without and  $P = 0.47$  with stratification based on the 11 groups), because most differences represented known (planned) options inherent to procedures. (3) Process variability was *not* associated with incidence of procedures ( $P = 0.79$ ), after excluding cataract surgery, a procedure with high relative variability. (4) Parameter uncertainty from uncommon procedures (0–2 historical cases) accounted for essentially all of the uncertainty in decisions dependent on estimates of OR times. The Bayesian method moderated the effect of small sample sizes on uncertainty in estimates of OR times. In contrast, from prior work, the use of broad categories of procedures reduces parameter uncertainty but at the expense of increased process variability.

**CONCLUSIONS:** For procedures with few historic data, the Bayesian method allows for effective case duration prediction, permitting use of detailed procedure descriptions. Although fine resolution of scheduling procedures increases the chance of performed procedure(s) differing from scheduled procedure(s), this does not increase process variability. Future studies need both to address differences in process variability among specialties and accept the limitation that findings from one may not apply to others. (Anesth Analg 2010;110:1155–63)

Predictive variability of operating room (OR) times influences decision making on the day of surgery including assignment of patients and nurses in the postanesthesia care unit, selection of start times of surgeons' add-on cases, assignment of relief staff, and deciding whether to move a case from one OR to another.<sup>1</sup> All of these decisions are affected by the uncertainty that is inherent in predictions of OR times, especially the time remaining in late running cases.<sup>2</sup>

Several factors contribute to predictive variability.

From the Departments of \*Anesthesia and Health Management and Policy, and †Management Sciences, University of Iowa, Iowa City, Iowa; and ‡Department of Thoracic Surgery, Roswell Park Cancer Institute, Buffalo, New York.

Accepted for publication December 31, 2009.

Franklin Dexter is the section Editor of Economics, Education, and Policy for the Journal. The manuscript was handled by Steve Shafer, Editor-in-Chief, and Dr. Dexter was not involved in any way with the editorial process or decision.

Address correspondence and reprint requests to Franklin Dexter, MD, PhD, Department of Anesthesia, University of Iowa, Iowa City, IA 52242. Address e-mail to Franklin.Dexter@UIowa.edu or Web site www.FranklinDexter.net.

Copyright © 2010 International Anesthesia Research Society

DOI: 10.1213/ANE.0b013e3181d3e79d

One contributor to predictive variability has been uncertainty in knowing the true probability distribution of OR times for a surgeon performing a procedure and how the probability distribution of OR times is centered about the scheduled case duration. However, the probability distributions of OR times based both on performed<sup>3,4</sup> and scheduled<sup>1,2,5,6</sup> procedures have been studied and are well understood<sup>7</sup> for single facilities.

A second contributor to variability in OR times is the process variability that arises among cases of the same procedure. For example, a surgeon has previously scheduled cataract extraction 350 times at the surgery center. Each such case has been scheduled to take 30 minutes. On average, the OR time is 33 minutes, with median absolute percentage error from the scheduled time of 35%. Such slight (3-minute) underestimation of OR time for many cases accounts epidemiologically for the majority of underestimation of OR times<sup>8</sup> and can be bias corrected for use when decisions are made on the day of surgery.<sup>9,10</sup> In contrast to the slight bias, the large 35% process variability substantively contributes to tardiness from scheduled start times of patients and surgeons using the same OR later on the same day.<sup>9–11</sup> One focus of our current study is the 35%

prediction error caused by process variability. Our objective is to better understand the process variability with the goal of its reduction.

A third contributor to predictive variability is parameter uncertainty caused by small sample sizes. Prediction intervals incorporate both process variability and uncertainty in the parameters. The smaller the sample size, the less accurate is the estimate of the mean and SD, and hence the wider the prediction interval. The width of a prediction interval can be expressed as the difference between the upper limit of the interval and the middle of the interval. For example, consider a surgeon who has previously scheduled anorectal myomectomy 2 times at a hospital. Both cases were scheduled for 2.5 hours of OR time. On the basis of just  $n = 2$  historical cases, the difference between the calculated 90% upper prediction bound for the OR time of the next case and the median prediction is much larger than what it would be if there was a larger sample size  $N$ . The difference between the 90% upper prediction bound for  $n = 2$  and the 90% upper prediction bound for a much larger  $N$  (e.g., an infinite amount of data) represents the parameter uncertainty. We focus on the upper tails (specifically 90%) of the predictive distribution, because many decisions on the day of surgery revolve around estimating the longest amount of time that an add-on case may take and/or on the expected time remaining for a case that has already taken longer than scheduled.<sup>1,2</sup>

Among all specialties at a facility, process variability is influenced mostly by procedure, but also by surgeon and type of anesthesia.<sup>12</sup> Regression models analyzing OR times after cases are completed have been developed for cardiac surgery<sup>13</sup> and adult endoscopy.<sup>14</sup> Among general thoracic cases, factors influencing process variability are, in sequence of importance: anatomic procedure, method and approach planned to achieve the anatomic result, surgical technique, composition of the surgical team, and type of anesthetic.<sup>15</sup> In our new study, *we test* (#1) whether process variability is homogeneous among specialties to learn whether insight from studying one specialty applies to other specialties. We use this knowledge in our companion article of a clinical trial designed to reduce variability.<sup>16</sup>

Previous epidemiological studies of process variability classified cases based on the procedures that were actually performed.<sup>12,15</sup> In contrast, studies designed for managerial applications rely on scheduled procedures, because when a case is scheduled, only the scheduled procedure is known.<sup>1,2,5,6</sup> This distinction is important not only scientifically<sup>7</sup> but also practically. Our impression is that many hospitals install their OR information systems to predict OR times from historical data classified by their *performed* procedure(s), perceiving that the distinction between scheduled and performed procedure(s) is unimportant and, if otherwise, surgeons should be more careful in selecting their scheduled procedures. *We test* (#2) whether change in procedure affects process variability by comparing a scheduled and actual procedure(s) classified using a common and systematic vocabulary (Current Procedural Terminology [CPT] codes). We hypothesize that when one procedure is planned but another is performed, the incidence of OR delays is increased, causing an increase in process variability. If true, the magnitude of the increase can guide

organizations in deciding whether to try to reduce predictive variability by focusing scheduling interventions on surgeons who often change procedures.

Previous studies revealed that many cases are of uncommon combinations of procedures. For example, in 1 study of 3 years of data, one-third of cases had only 0 to 2 other cases of the same combination of surgeon, scheduled procedure(s), and type of anesthetic.<sup>6</sup> More than half of the cases were of a procedure scheduled by the surgeon <3 times per year.<sup>17</sup> Although the broad category of procedure is the same (e.g., "coronary artery bypass grafting"), the precise procedure(s) differs (e.g., vessel harvesting sites) and has different OR times (i.e., the use of broad categories increases process variability). Pooling data among facilities can be ineffective to increase sample sizes substantively, because a procedure(s) that is uncommon at 1 facility tends to be uncommon elsewhere.<sup>18</sup> Among outpatient cases in the United States with an anesthesia provider, 20% were of procedures performed  $\leq 4$  times per workday *nationwide* and 36% of cases were of procedures performed an average of <1 time per facility per year.<sup>19</sup> These results are not surprising, because community hospitals have reported >5600 surgeon preference cards<sup>20</sup> and academic hospitals >13,000 such cards.<sup>21</sup> Previously, we found no correlation between incidence of use of surgeon preference cards\* and process variability as measured by absolute percentage error.<sup>15</sup> In other words, uncommon procedures likely were reducing sample sizes  $N$ , but not increasing SDs of OR times for each surgeon preference card. In our new study, we perform the same assessment but use data from a different hospital that scheduled using the systematic vocabulary of CPT codes, rather than a local system of preference cards. *We test* (#3) whether facilities that are focused on performing only a few high-volume procedures can expect less process variability than facilities for which many cases are uncommon procedures.<sup>22</sup> This is an important question because it contributes to previous studies<sup>9</sup> of the value (or not) of a "focused factory" model for surgery.

Finally, the principal influence of uncommon procedures on uncertainty in decision making on the day of surgery may be not only via process variability but also parameter uncertainty from small sample sizes. In our new study, we explore the relative importance of process variability and parameter uncertainty on the predictive variability. *We test* (#4) the extent to which a Bayesian method,<sup>2,6,8</sup> implemented for ongoing updates of case duration predictions, serves to mitigate the impact of parameter uncertainty on the predictive variability of OR times. The answer is important to understanding what techniques need to be implemented by facilities to achieve substantive improvements in the quality of its managerial decisions on the day of surgery.

## METHODS

### Analysis of the Dataset Used in the Forthcoming Results Sections #1, #2, and #3

We studied all 61,353 cases performed at a hospital's tertiary surgical suite ( $n = 43,614$ ) or outpatient surgery

\*Specified surgical trays and instruments, positioning and equipment, drains, tubes, and catheters.

**Table 1. Group Assigned to Each Case**

Description of group assigned	Surgeon's most common Clinical Classifications Software (CCS) code	Elective or urgent	Surgical suite	N	Scheduled OR time (h) (mean ± sb)	Abbreviation in Figure 1
Joint replacement	Hip replacement, total and partial (153)			451	3.2 ± 0.8	Joint
Elective spine	Laminectomy, excision intervertebral disc (3), or Spinal fusion (158)	Elective		3280	4.7 ± 2.3	Spine elec
Urgent spine	Laminectomy, excision intervertebral disc (3), or Spinal fusion (158)	Urgent		1243	3.4 ± 1.4	Spine urg
Cardiac	Coronary artery bypass graft (44)			1959	5.4 ± 2.4	Cardiac
Urology	Nephrectomy; partial or complete (104)			615	3.6 ± 1.9	Urological
Gynecology	Hysterectomy; abdominal and vaginal (124)			1152	3.0 ± 1.2	Gynecology
Colorectal	Colorectal resection (78)			171	3.4 ± 1.4	Colorectal
General thoracic	Lobectomy or pneumonectomy (36)			265	4.1 ± 2.0	Thoracic
Elective case at tertiary site	Any of the other 178 CCS with at least 1 case	Elective	Tertiary suite	26,231	3.2 ± 1.8	Main elec
Ambulatory center	Any of the other 178 CCS with at least 1 case		Ambulatory center	17,739	1.7 ± 0.9	ASC
Urgent case at tertiary site	Any of the other 178 CCS with at least 1 case	Urgent	Tertiary suite	8247	3.0 ± 1.6	Main urg

The 11 groups are listed as closely to the sequence shown in Figure 1 as possible based on the definitions used.

center ( $n = 17,739$ ) between June 1, 2002 and May 31, 2005. The time from “wheels in” to “wheels out” was considered the “OR time.” The scheduled OR time was provided by the surgeon and/or scheduler. For each case, there were 1 to 5 scheduled and 1 to 5 actual CPTs. The CPTs were considered without regard to sequence, because the sequence was arbitrary. For example, it does not matter whether a case is scheduled as “adenoidectomy and myringotomy” or “myringotomy and adenoidectomy.”

The scheduled OR times were slightly biased, with the mean overestimate of OR time being 10.7 minutes per case. Although minor relative to the mean OR time of 163 minutes, for the scientific analysis, we needed to reduce the bias (see Discussion section). We corrected the bias by using linear least squares regression of actual and scheduled OR time, with coefficients estimated after exclusion of the 1.8% of cases with absolute Studentized residuals  $>3$  (SYSTAT 12, Systat Software, San Jose, CA). All cases were used for all analyses other than the estimation of the 2 regression parameters. Bias correction such as this is applied once cases have been assigned to ORs and the schedule is being finalized, for use on the day of surgery.<sup>2,10</sup> The bias-corrected scheduled OR time =  $-4.4$  minutes +  $0.946 \times$  scheduled OR time in minutes, with standard errors of the coefficients of 0.4 minutes and 0.002, respectively. The mean  $\pm$  SD of the absolute difference between the original and bias-corrected scheduled times was  $13 \pm 5$  minutes. The Pearson correlation coefficient between the actual and scheduled OR times was  $r = 0.858$ , with SE 0.001. We used linear regression because it allowed us to correct (from 10.7 minutes down to 2.7 minutes) for the relatively small bias with a parsimonious model consisting of only 2 parameters, as compared with the enormous sample size of  $n = 61,353$ . Correcting the scheduled OR time individually for each procedure would not have achieved that objective because there were 17,571 different procedures.

Process variability was quantified (below) for each case by using the absolute percentage error calculated as 100 times the absolute value of 1 minus the ratio of the

bias-corrected scheduled OR time to the actual OR time. Within each group described below, the mean absolute error from each case's scheduled OR time would be a suitable end point,<sup>16</sup> because direct economic cost (e.g., to anesthesiologists) and indirect/intangible cost (e.g., to surgeons following a colleague in the same OR on the same day) are related to absolute errors. However, to compare among groups, the mean absolute percentage error was used because of large differences among groups in mean scheduled OR times (e.g., see Table 1). We quantified process variability using a measure of relative error because by far the single largest predictor of absolute error is the scheduled duration (i.e., a 30-minute cataract extraction may finish 5 minutes late whereas a 6-hour esophagectomy may finish 1 hour late). This is already known and trivially incorporated into existing methods of case duration prediction.<sup>1-12</sup>

### Section #1

We created groups of procedures based on our expectations for current and future retrospective and prospective clinical trials of case duration prediction (Table 1). For example, retrospective studies have been performed for ambulatory surgery,<sup>9,10</sup> cardiac surgery,<sup>13</sup> general thoracic surgery,<sup>15</sup> and joint replacement surgery.<sup>23</sup> Thus, we included these groups. In our companion article, we prospectively studied general thoracic surgery and elective spine surgery.<sup>16</sup> For these, we created 4 of the groups in Table 1. The groups were created from among specialties that could be defined based on their International Classification of Diseases procedure codes and CPT codes.<sup>24</sup> From the 1 to 5 CPT codes of each case, we obtained 1 to 5 Clinical Classifications Software (CCS) categories. Each case's specialty was chosen based on the surgeon's most common CCS category.<sup>24</sup> The CCS categories were accessed from [http://www.hcup-us.ahrq.gov/toolssoftware/ccs\\_svcsproc/ccssvcproc.jsp](http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp) on April 1, 2009. A case was considered to be “elective” if it was scheduled at least 1 workday before the day of surgery, because scheduled cases are performed on workdays.<sup>1,7,9,10</sup> This functional definition matches that which can be used in prospective

studies to reduce predictive variability of OR times. Other cases were called “urgent” (Table 1).

The Pearson  $\chi^2$  test was used to compare the numbers of cases in each of the 11 groups for which the absolute percentage error exceeded or did not exceed the overall ( $n = 61,353$ ) median of 22%. This happens to be the so-called “median test” for differences among the 11 groups. The  $\chi^2$  median test was also used to compare each group with the overall 90th percentile of the absolute percentage error. Effect size for the  $2 \times 11$  contingency tables was quantified by using Cramer’s V, with confidence intervals (CIs) calculated asymptotically (StatXact-8, Cytel Software Corporation, Cambridge, MA). In other words, Cramer’s V was used as a post hoc test to determine the strength of association after the  $\chi^2$  test had determined its statistical significance.

### Section #2

Each case’s combination of 1 to 5 scheduled CPTs was compared with the case’s combination of 1 to 5 final CPTs. Each case’s absolute percentage error in scheduled versus actual OR time was classified as exceeding or not exceeding the overall median of 22%. A  $2 \times 2$  contingency table was created containing the numbers of cases with absolute percentage errors (a) exceeding or (b) not exceeding the median among cases with (c) all CPTs the same and (d) 1 or more CPTs different. The table was analyzed by using the Pearson  $\chi^2$  test. The analysis was repeated after stratifying the  $2 \times 2$  contingency tables into the 11 groups of Table 1. Those calculations were performed using the Mantel-Haenszel test (StatXact-8).

In section #2 of the Results, we show that although changes in procedure(s) modestly increase OR times, absolute percentage errors are not increased. To explore why this occurs, we focused on 2 specific subgroups based on the CCS compilation of similar CPTs. In our studied data, “skin graft” (CCS 172) was the most common CCS category for which the scheduled combination of CPTs did not equal the actual CPTs. Among those 2052 cases, 94% of cases had at least 1 difference between scheduled and final CPTs. Our methodology was to repeat the analyses of the preceding paragraph for those cases. Next, among cases performed in the ambulatory surgery center, “lens and cataract procedures” (CCS 15) was the most common CCS category for which the scheduled combination of CPTs did not equal the actual CPTs. Those 3134 cases had 19% with a change in at least 1 CPT. We repeated the analyses of the preceding paragraph for those cases as well.

### Section #3

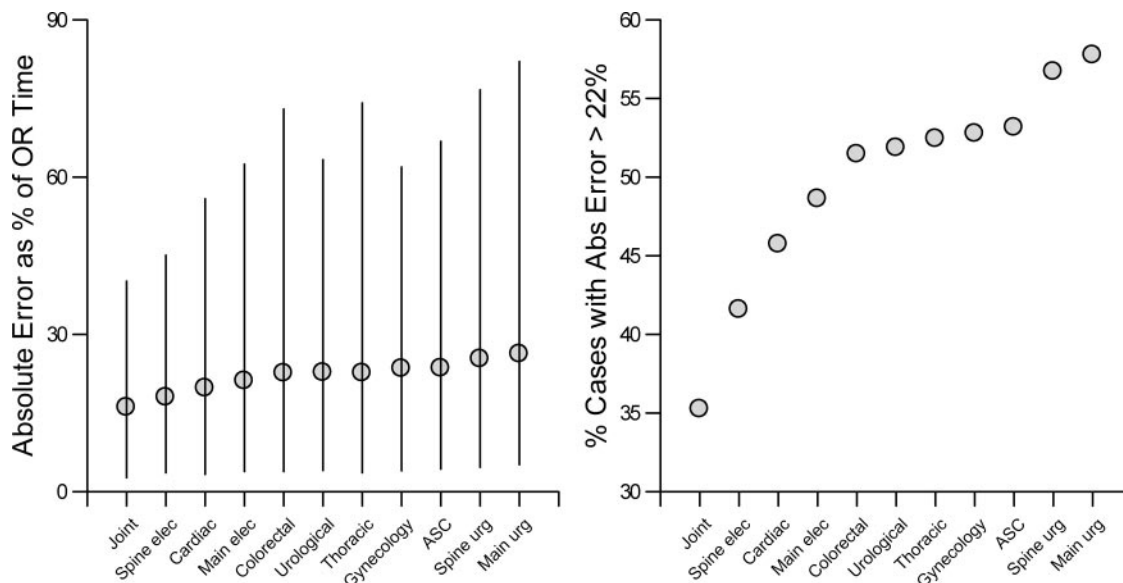
We tested for a relationship between incidence ( $N$ ) of each combination of CPTs and the absolute percentage error. Incidences (e.g.,  $n = 1$  and  $n = 12$ ) were divided into categories (e.g.,  $n = 1$  and  $9 \leq N \leq 18$ ) and each included several different numbers of cases. The categories could not be chosen so that each category included precisely 1 decile of all cases, because the incidences are discrete values. Consequently, thresholds defining the smallest and largest numbers of cases for each category were chosen so that each category included as close to deciles of the population as possible given that the numbers of cases were integers.

The resulting cut-points were as follows:  $n = 1$  (cumulative 22%),  $n = 2$  (27%),  $3 \leq N \leq 8$  (40%),  $9 \leq N \leq 18$  (50%),  $19 \leq N \leq 36$  (60%),  $37 \leq N \leq 70$  (70%),  $71 \leq N \leq 129$  (80%),  $130 \leq N \leq 299$  (90%), and  $N \geq 300$  (100%). These incidences refer to the  $N$  for the studied 3-year period (i.e., when  $n = 1$ , there would have been 0 historical data on which to make a decision involving OR times). The Cochran-Armitage trend test was used to test for association between decile of frequency and numbers of cases with percentage absolute error exceeding the overall median (StatXact-8). The 2-sided trend test compares the null hypothesis of equal percentages with the omnibus alternative hypothesis of  $p_{n=1} \leq p_{n=2} \leq p_3 \leq N \leq 8 \leq \dots \leq p_{N \geq 300}$ , or vice versa, with the former consistent with the hypothesis of decreasing variability for increasing incidence.

### Analysis of the Dataset to Investigate Parameter Uncertainty in Results Section #4

Regardless of whether incidence influences (or does not influence)<sup>15</sup> uncertainty in decisions involving OR times by causing a change in process variability (above), incidence influences decision making on the day of surgery via parameter uncertainty.<sup>1,2,5,6</sup> One way to quantify and intuitively understand parameter uncertainty is to examine the width of the prediction interval expressed by the ratio of the 90% prediction bound to the 50% quantile of a future case.<sup>2,6</sup> The ratio is principally a function of the process variability in the data (e.g., SD) and the amount of data (sample size,  $N$ ) on which predictions are based, the latter affecting the cutoff of the Student  $t$  distribution.<sup>1,5</sup> When there are  $n = 200$  historical data, the 90% prediction bound is 1.3 SD from the median.<sup>1,5</sup> When there are  $n = 2$  historical data, the 90% prediction bound is 3.8 SD from the median.<sup>1,5</sup> For the 27% of cases with  $n = 0$  or  $n = 1$  other cases, the 90th percentile of the Student  $t$  statistic is essentially infinite. For example, suppose that at 10 AM there is an add-on case scheduled for 1.5 hours and the surgeon is available. Including 30-minute turnover time, the scheduled time needed would be 2.0 hours. A suitable OR is idle until 12 noon, at which time another surgeon who consistently shows up on time is scheduled to start a list of 6 short cases. Staffing is planned for that OR until 3 PM, with the last of the 6 cases scheduled to end at 2:45 PM. The decision of whether to use the OR for the add-on case depends on the risk of delaying the to-follow surgeon and his patients and of delaying the end of the workday.<sup>2,5</sup> The risk of causing delay is high if that add-on case is of a procedure that the surgeon has not previously performed at the facility (i.e.,  $n = 0$ ).<sup>1,2,5,6</sup>

Bayesian methods moderate the influence of small sample sizes on parameter uncertainty.<sup>2,6</sup> When a case is of a procedure(s) with few or no historical data, the previously developed Bayesian method<sup>2,6</sup> relies on the scheduled OR time as the principal or sole basis for future prediction of the median OR time. The proportional predictive variability of the case is estimated from the cases with many historical data. This is somewhat similar to assuming that all of the cases with few or no historical data share the same median absolute percentage error (see Fig. 8 of Ref. 6). Thus, the 90% prediction bound is estimated using both the scheduled OR time (with bias correction if



### Each Case into One Group

**Figure 1.** Differences in predictive variability among groups. The overall 10th, 50th, and 90th percentiles of absolute percentage errors were 4.2%, 22.3%, and 65.8%, respectively. The prediction is the surgeon scheduled operating room (OR) time, with correction for the slight bias present in the estimates (see second paragraph of Methods section). The statistical analysis in the Results section analyzed the numbers of cases in each group with absolute percentage errors exceeding these thresholds. The panel on the right gives an example of results for the 50th percentile (i.e., for the Median Test). On the vertical axis on the right, “Abs Error” refers to absolute percentage error.

persistently underestimating or overestimating)<sup>2</sup> and the proportional predictive variability from cases of many different procedures. In contrast, when a case has substantial historical data, effectively no assumption is being made about the absolute percentage error. The scheduled OR time has a negligible effect on the Bayesian predictions compared with the information from the historical OR times of the surgeon and scheduled procedure(s).<sup>2,6,8</sup> The magnitude of the effect that the Bayesian method has on mitigating the influence of parameter uncertainty on uncertainty in OR times is unknown and is the focus of our final investigation (#4).

The data studied were from the same hospital as the preceding dataset, but the dates are different. The dates studied were those used previously for parameter estimation and validation of the Bayesian method, so that we do not need<sup>8</sup> to repeat the relevant analysis from earlier publications (including several pages of notation, equations, methodology, and results).<sup>2,6</sup> Specifically, we showed previously that the Bayesian 90% prediction bounds used in our current article are exceeded by the actual duration of 9.7% of cases, close to the expected 10% value.<sup>6</sup> We could not use these same data for studies 1, 2, and 3, because both scheduled and actual CPTs were not available in those data.

The 65,661 cases were performed between January 1, 1996 and December 31, 1999.<sup>6</sup> For each case, calculations relied on all other cases of the same combination of surgeon, scheduled procedure(s), and presence or absence of an anesthesia provider.<sup>2,12,6</sup> There were 19,838 such combinations and each case had 0 to 917 other cases of the same combination. Referring to equation numbers from the original article,<sup>6</sup> 90% prediction bounds were

calculated with the Bayesian method’s equation (1) and the median with equation (2). Starting with 0 other cases, we created cumulative categories with at least 2000 cases. We created categories in that manner instead of using deciles, because we wanted to focus disproportionately on uncommon combinations of surgeon and scheduled procedure(s).

## RESULTS

### 1. Magnitude of Process Variability and Differences Among Groups

The 50th percentile of all absolute percentage errors was 22%. The 90th percentile was 66%.

Groups differed significantly in the distributions of cases above and below the 50th and 90th percentiles (both  $P < 0.00001$ , Fig. 1). For example, 1 group had only 42% of absolute percentage errors exceeding the overall median of 22%. That group of elective spine surgery had the second smallest median absolute percentage error. In comparison, urgent spine surgery had the second largest median absolute percentage error, exceeding the overall median of 22% for 57% of cases.

The magnitude of the differences among the groups did not differ significantly between the 50th and 90th percentiles (Cramer’s  $V$ , which ranges from 0 to 1, equaled 0.085 [95% CI, 0.077–0.093] for the 50th percentile vs 0.085 [95% CI, 0.078–0.093] for the 90th percentile). Thus, we focus on the median absolute percentage errors.

### 2. Influence of Scheduled Versus Actual CPTs on Process Variability

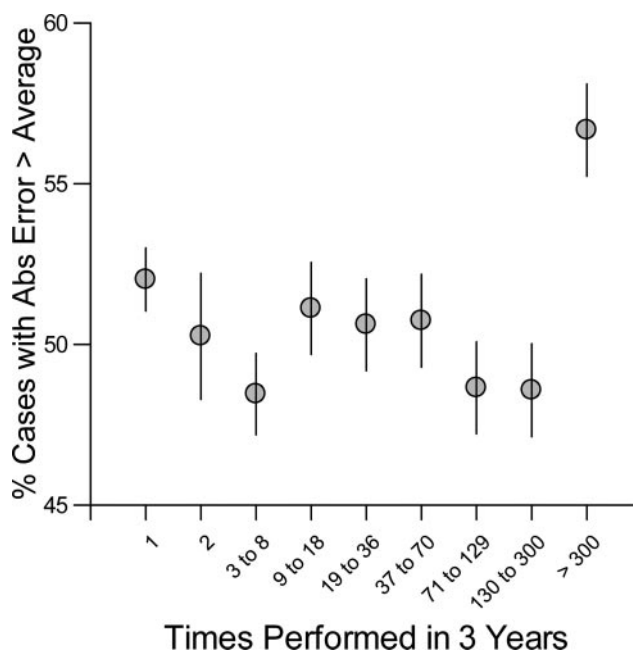
Absolute percentage errors were the same regardless of whether the scheduled and actual CPTs were the same or

different. The overall 50th percentile was exceeded by 50% of cases for both scenarios ( $P = 0.23$  without and  $P = 0.47$  with stratification based on the 11 groups). The 95% CIs are 50% to 51% for same and 50% to 52% for different CPT codes.

We had expected results such as the finding that joint replacement was both the group with the largest percentage of cases with all CPTs the same (75%) and the group with the smallest process variability (Fig. 1). However, even for general thoracic surgery with its large (Fig. 1) and well-studied<sup>15</sup> process variability, there was no significant difference in the percentages of cases with absolute percentage errors exceeding the overall median among cases with unchanged versus changed CPTs (56% vs 52%,  $P = 0.56$ ).

Partially, the cause reflects the skewness of the probability distributions of OR times. If the duration of staffing for each OR is chosen appropriately months before individual cases are scheduled, each case should be scheduled into the OR time using an unbiased estimator for the case's contribution to the total workload that was used to plan the staffing.<sup>6,8,26,27</sup> Because of a slight right skewed probability distribution, more than half of cases take less time than the mean.<sup>2,6,8</sup> The value is 56% for the studied cases (95% CI, 56%–57%). Because 56% of cases take less time than the mean, the estimate that minimizes the bias differs slightly from the estimate that minimizes the process variability. When the actual CPTs differ from the scheduled CPTs, a smaller percentage of cases (51%) take longer than scheduled (95% CI, 50%–52%). Thus, although a change in procedure(s) slightly lengthens the case duration, the effect is a reduction in the absolute percentage error. This statistical nuance explains why the absolute percentage errors did not differ even though changing procedures added an average of 9 minutes to cases. Regardless, the magnitudes of these differences are too small to be of managerial importance, provided they are accounted for when cases' scheduled start times are chosen.<sup>8–10</sup>

We used selected CCS to understand why changing procedures did not usually result in the large increases in OR time that we had expected. The most common CCS category for which the scheduled and actual combinations of CPTs differed was "skin graft" (see Methods section). Many of those CPTs specify locations and square centimeters autografted, transferred, etc. Changes in CPTs represented expected decision options that the burn surgeons made intraoperatively. Repeating the analysis for the ambulatory surgery center cases only, the most common CCS category for which the CPT(s) changed was "lens and cataract procedures." More than half of the scheduled CPTs not among final CPTs were 66984, "extracapsular cataract removal with insertion of intraocular lens prosthesis (1-stage procedure), manual or mechanical technique (e.g., irrigation and aspiration or phacoemulsification)." More than half of the final CPTs not among the scheduled CPTs were CPT code 66982. The latter is the same as 66984 with the addition: "complex, requiring devices or techniques not generally used in routine cataract surgery (e.g., iris expansion device)." Thus, most differences between scheduled and final CPTs represented known (planned) options inherent to the surgery.



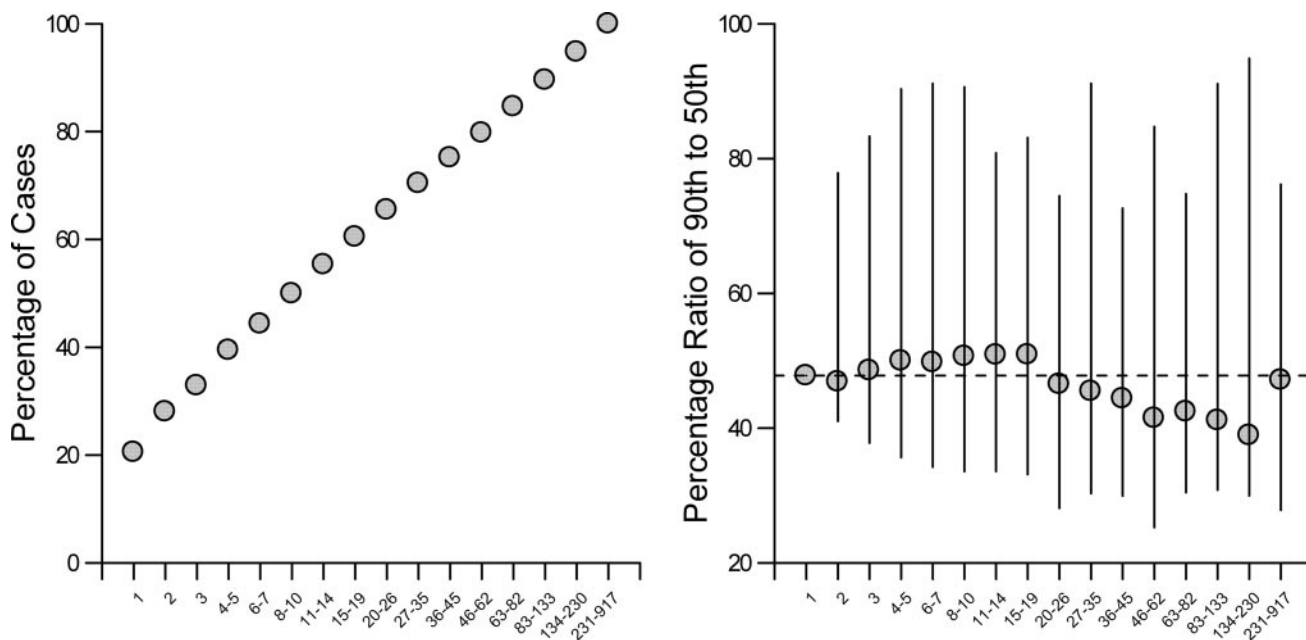
**Figure 2.** Influence of incidence of procedures on predictive variability. The absolute percentage error was calculated for each case. The percentage of cases exceeding the median absolute percentage error is plotted for each incidence category, corresponding to how the median test works. The upper and lower bars show 95% confidence intervals deliberately calculated *without* correction for multiple comparisons (i.e., they underestimate the actual widths that would maintain the family-wise error rate). Increasing incidence of procedures is not associated ( $P = 0.79$ ) with ordered differences in predictive variability excluding the category of >300 cases, as described in section #3 of the Results section.

### 3. Uncommon Versus Common Combinations of Procedures

Increasing incidence of procedures was associated with *larger* process variability ( $P < 0.00001$ , Fig. 2). However, this was attributable solely to the decile of the most common procedures, 57% of which had absolute percentage errors > median. All cases of the uppermost decile had just 1 performed CPT. The most common CPT code among cases of this decile was 66984, described in the preceding paragraph. These cataract surgery cases had substantial relative process variability, with 82% of cases having absolute percentage error > median. Excluding this decile, there was no relationship between incidence and process variability ( $P = 0.79$ ), with observed percentages of cases with absolute percentage error exceeding the median ranging only from 48% of cases to 52% of cases. The latter finding matched the results of our prior work from a different hospital using surgeon preference cards.<sup>15</sup>

### 4. Parameter Uncertainty

The ratio of the 90% prediction bound to the 50% quantile for a case is principally a function of the SD and the 90th percentile of the Student *t* statistic (see Introduction and Methods sections).<sup>1,5</sup> For the 27% of cases with  $n = 0$  or  $n = 1$  other cases, the *t* statistic is essentially infinite, and thus so is the resulting ratio. The footnote below shows that, from principles of limits, those cases are responsible for essentially all of the uncertainty in decisions dependent on



## Cases of Same Surgeon and Scheduled Procedure(s)

**Figure 3.** Influence of Bayesian method on parameter uncertainty. The panel on the left shows the cumulative probability distribution for the number of cases of the same combination of surgeon, scheduled procedure(s), and scheduled type of anesthesia. Each range has at least 2000 cases (see Methods section). The result of importance is that one-third of the cases had 0 to 2 other similar cases. The panel on the right shows the ratio of the 90% to the 50% prediction bounds from the Bayesian methods. The large heights of the bars from the 10th to 90th percentiles show that there is much heterogeneity in ratios among surgeons and procedures, reflecting different process variability.<sup>6</sup> The implication is that decisions need to rely on historical data for each surgeon and procedure.<sup>6</sup> The circles show the median. For 0 historical data, all 13,451 cases have the same ratio, because that ratio is estimated entirely from data of other combinations of surgeon and procedures. As explained in the Results section, the important observation is that when the Bayesian method is applied, the median ratios are practically the same. Without the Bayesian method, the ratios are essentially infinite for  $n = 1$  and  $n = 2$ ,† and several-fold larger for  $n = 3$ .

estimates of OR times.† The limiting argument is extreme, of course, and OR managers knowing that it is not infinite apply contextual (prior) knowledge (i.e., qualitatively use a Bayesian method). Nevertheless, the results underestimate the actual uncertainty in decisions because most managerial decisions on the day of surgery that involve predictions of OR times involve comparing the OR times of multiple cases.<sup>1,28-33</sup> If any one such case involved in the decision has 0 or 1 historical data, then the decision making is affected by parameter uncertainty.

The ratio of the 90% Bayesian prediction bound to the Bayesian median was calculated for each case. Each vertical line in Figure 3 shows the 10th percentile, median (indicated by the open circle), and 90th percentile of these ratios. The 3 end points of the ratios differ little among sample sizes. Thus, Figure 3 shows that use of the Bayesian method results in near complete moderation of the effect of small sample sizes on uncertainty in estimates of OR times.

## DISCUSSION

We performed our study to better understand the magnitude of the influence of how surgical procedures are

classified on both process variability and parameter uncertainty and what to do about the relationships. We knew a priori that there would be relationships because there are large differences in OR times depending on the anatomic procedure used for the same medical condition and on the surgical approach used to achieve the anatomic result.<sup>15</sup> Using an extreme example, if broad categories such as the CCS were used, there would be virtually no parameter uncertainty, but larger process variability.<sup>2,6</sup>

It was already known that the one-third of cases with no or few historical data account for uncertainty in decisions involving prediction of OR times.<sup>1,5,6</sup> The new results revealed that: (#1) those cases accounted for most (if not even relatively all) of the uncertainty; (#2, #3) the cause was almost entirely parameter uncertainty because process variability does differ based on sample size; and (#4) the Bayesian method nearly fully moderated the effect of parameter uncertainty. The benefits would be accrued when making decisions on the day of surgery, at which time many if not most decisions involving OR times depend on the shortest and longest times cases may take and on the times remaining in long running cases (i.e., on the tails of the OR time probability distributions).<sup>1,2,5,6</sup> The implication of our findings is that no substantive improvements in decision making on the day of surgery at tertiary (multidisciplinary) facilities

†Let  $x$  represent contribution from cases with 0 or 1 historical data and  $r$  represent the remainder. The proportional effect of the former equals the limit as  $x \rightarrow \infty$  of  $x/(x+r)$ . Letting  $y = 1/x$  and substituting into the ratio, the limit is the same as  $y \rightarrow 0$  of  $1/(1+ry) = 1$ .

can be expected at facilities without use of the Bayesian method<sup>2,6</sup> or equivalent<sup>34</sup>.

It was already known from studies of the Bayesian method at 2 hospitals<sup>2,6</sup> that data for each procedure should be used whenever possible because process variability differs among procedures. The new results (section #1) revealed that process variability differed among procedures at least partly because of systematic differences among specialties. This knowledge is important for performing and interpreting observational studies of case duration variability.<sup>15,35</sup> In our companion article, we describe our clinical trial using 2 specialties chosen deliberately based on their having widely different process variabilities (Table 1).<sup>16</sup>

Finally, it was already known that for unbiased managerial decision making, historical OR time data need to be analyzed (e.g., with Bayesian method) based on scheduled procedure(s), not actual procedure(s), because scheduled procedures are known when future decisions are made.<sup>36,37</sup> The new results revealed that: (section #2) changes in procedure(s) were frequent when a systematic vocabulary was used, (#2) the changes were often expected (planned) intraoperative options, and (#3) systematic reduction in the incidence of changes would not substantively reduce overall process variability. Classifying cases based on scheduled versus actual procedure(s) results in biased estimates<sup>36,37</sup> but (#2) overall does not affect relative process variability. Because many interventions would need to be made to prevent a few marked delays, the implication is that interventions targeting change in procedure(s) as an adverse managerial event are likely to be frustrating to surgeons and without substantive overall benefit from the perspective of case duration prediction and associated managerial decisions.

Although our finding that short ambulatory cases have relatively large relative variabilities was insightful mechanistically (see last paragraph of Results section #2), we suspect that it is unimportant managerially. The reason is that process variability influences decision making on the day of surgery mostly for the coordination of patient flow from OR to postanesthesia care unit, selection of start times of surgeons' add-on cases, decision of whether to move a case from one OR to another OR, and assignment of relief staff.<sup>1</sup> These decisions tend to be unimportant at ambulatory surgery centers.<sup>38</sup> How long patients wait and are fasting on the day of surgery are important,<sup>38</sup> but such decisions generally need to be communicated before there are complete data on all preceding cases in an OR, and thus other statistical methods apply.<sup>7</sup> How long staff members work including overtime matters, but if decisions are made optimally months before the day of surgery, then decisions are influenced little by typical magnitudes of predictive variability in case durations.<sup>1,8,26,39</sup> The reason is that surgery is a non-preemptive task and the period of largest variability is the portion that is non-preemptive (unlike in other industries such as airlines). Thus, staffing decisions made months in advance should incorporate and mitigate much of the influence of the predictive variability.<sup>1,8,26,39</sup>

Although we had to consider bias in scheduled OR times for our study of imprecision to be scientifically valuable (i.e., applicable to other facilities), bias is generally

unimportant as long as it is considered and compensated for statistically when decisions are made.<sup>2,9,10,40</sup> For example, conceptually it would seem fruitful to implement a protocol not to rely on scheduled estimates that differ significantly ( $P < 0.05$ ) from the historical average OR times of cases of the same surgeon and scheduled procedure(s).<sup>2,8</sup> However, we showed previously that this intervention is not of value because for cases with a sufficient number of historical data for a small  $P$  value to be calculated, predictive variability in OR times results principally from process variability, not bias.<sup>2,8</sup>

From knowledge of the probability distributions of OR times,<sup>5,12</sup> the Bayesian approach was developed to estimate OR times regardless of whether the case is of a combination of surgeon and procedure(s) with no, a few, or many historical cases of the same procedure(s).<sup>6</sup> Implementation for use on the day of surgery can be done without human data entry by automatically inferring case progress from networked vital signs or from the anesthesia information management system.<sup>2,41,42</sup> Figure 1 suggests the potential for improvement in our formulation of the Bayesian method. Currently, for few historical data, the method uses all cases to estimate relative process variability.<sup>6</sup> Figure 1 suggests that there might be value in using information by specialty. The statistical methodology has already been described.<sup>6,35</sup> However, we do not yet know the reasons for the differences among specialties and the resulting stability in them over time. What the current study's results have shown is that the cause is neither change in procedure nor differences in sample size. Our companion article focuses on this area. ■■

## REFERENCES

- Dexter F, Epstein RD, Traub RD, Xiao Y. Making management decisions on the day of surgery based on operating room efficiency and patient waiting times. *Anesthesiology* 2004;101:1444–53
- Dexter F, Epstein RH, Lee JD, Ledolter J. Automatic updating of times remaining in surgical cases using Bayesian analysis of historical case duration data and instant messaging updates from anesthesia providers. *Anesth Analg* 2009;108:929–40
- Strum DP, May JH, Sampson AR, Vargas LG, Spangler WE. Estimating times of surgeries with two component procedures: comparison of the lognormal and normal models. *Anesthesiology* 2003;98:232–40
- Spangler WE, Strum DP, Vargas LG, May JM. Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health Care Manag Sci* 2004;7:97–104
- Zhou J, Dexter F. Method to assist in the scheduling of add-on surgical cases: upper prediction bounds for surgical case durations based on the log normal distribution. *Anesthesiology* 1998;89:1228–32
- Dexter F, Ledolter J. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historical data. *Anesthesiology* 2005;103:1259–67
- Wachtel RE, Dexter F. Simple method for deciding what time patients should be ready on the day of surgery without procedure-specific data. *Anesth Analg* 2007;105:127–40
- Dexter F, Macario A, Ledolter J. Identification of systematic under-estimation (bias) of case durations during case scheduling would not markedly reduce over-utilized operating room time. *J Clin Anesth* 2007;19:198–203
- Wachtel RE, Dexter F. Influence of the operating room schedule on tardiness from scheduled start times. *Anesth Analg* 2009;108:1889–901

10. Wachtel RE, Dexter F. Reducing tardiness from scheduled start times by making adjustments to the operating room schedule. *Anesth Analg* 2009;108:1902–9
11. Denton B, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Manag Sci* 2007;10:13–24
12. Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* 2000;92:1454–67
13. Lehtonen JM, Kujala J, Kouri J, Hippelainen M. Cardiac surgery productivity and throughput improvements. *Int J Health Care Qual Assur* 2007;20:40–52
14. Combes C, Meskens N, Rivat C, Vandamme JP. Using a KDD process to forecast the duration of surgery. *Int J Prod Econ* 2008;112:279–93
15. Dexter F, Dexter EU, Masursky D, Nussmeier NA. Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesth Analg* 2008;106:1232–41
16. Dexter EU, Dexter F, Masursky D, Kasprovicz K. Prospective trial of thoracic and spine surgeons' updating of their estimated case durations at the start of cases. *Anesth Analg* 2010;110:1164–8
17. Zhou J, Dexter F, Macario A, Lubarsky DA. Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. *J Clin Anesth* 1999;11:601–5
18. Dexter F, Traub RD, Fleisher LA, Rock P. What sample sizes are required for pooling surgical case durations among facilities to decrease the incidence of procedures with little historical data? *Anesthesiology* 2002;96:1230–6
19. Dexter F, Macario A. What is the relative frequency of uncommon ambulatory surgery procedures in the United States with an anesthesia provider? *Anesth Analg* 2000;90:1343–7
20. Wortmann KD, Weeks T. Successful: preference list building. *Surg Serv Manag* 1998;4(11):20–6
21. Porter J. Building doctor's preference cards. *Surg Serv Manag* 1999;5(5):43–7
22. Wachtel RE, Dexter F. Differentiating among hospitals performing physiologically complex operative procedures in the elderly. *Anesthesiology* 2004;100:1552–61
23. Dexter F, Weih LS, Gustafson RK, Stegura LF, Oldenkamp MJ, Wachtel RE. Observational study of operating room times for knee and hip replacement surgery at nine US community hospitals. *Health Care Manag Sci* 2006;9:325–39
24. O'Neill L, Dexter F. Tactical increases in operating room block time based on financial data and market growth estimates from data envelopment analysis. *Anesth Analg* 2007;104:355–68
25. Stepaniak PS, Mannaerts GH, de Quelerij M, de Vries G. The effect of the operating room coordinator's risk appreciation on operating room efficiency. *Anesth Analg* 2009;108:1249–56
26. Dexter F, Macario A, Traub RD. Which algorithm for scheduling add-on elective cases maximizes operating room utilization? Use of bin packing algorithms and fuzzy constraints in operating room management. *Anesthesiology* 1999;91:1491–500
27. Dexter F, Traub RD. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesth Analg* 2002;94:933–42
28. Dexter F, Traub RD, Qian F. Comparison of statistical methods to predict the time to complete a series of surgical cases. *J Clin Monit Comput* 1999;15:45–51
29. Dexter F, Traub RD. Sequencing cases in operating rooms—predicting whether one surgical case will last longer than another. *Anesth Analg* 2000;90:975–9
30. Dexter F, Traub RD. Statistical method for predicting when patients should be ready on the day of surgery. *Anesthesiology* 2000;93:1107–14
31. Dexter F, Traub RD, Lebowitz P. Scheduling a delay between different surgeons' cases in the same operating room on the same day using upper prediction bounds for case durations. *Anesth Analg* 2001;92:943–6
32. Dexter F, Willemssen-Dunlap A, Lee JD. Operating room managerial decision-making on the day of surgery with and without computer recommendations and status displays. *Anesth Analg* 2007;105:419–29
33. Dexter F, Lee JD, Dow AJ, Lubarsky DA. A psychological basis for anesthesiologists' operating room managerial decision-making on the day of surgery. *Anesth Analg* 2007;105:430–4
34. Stepaniak PS, Heij C, Mannaerts GH, de Quelerij M, de Vries G. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesth Analg* 2009;109:1232–45
35. Olivares M, Terweisch C, Cassorla L. Structural estimation of the newsvendor model: an application to reserving operating room time. *Manage Sci* 2008;54:41–55
36. Dexter F. Application of prediction levels to operating room scheduling. *AORN J* 1996;63:607–15
37. Macario A. Truth in scheduling: is it possible to accurately predict how long a surgical case will last? *Anesth Analg* 2009;108:681–5
38. Smallman B, Dexter F. Optimizing the arrival, waiting, and NPO times of children on the day of pediatric endoscopy procedures. *Anesth Analg* 2010;110:879–87
39. McIntosh C, Dexter F, Epstein RH. Impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: tutorial using data from an Australian hospital. *Anesth Analg* 2006;103:1499–516
40. Pandit JJ, Dexter F. Lack of sensitivity of staffing for 8 hour sessions to standard deviation in daily actual hours of operating room time used for surgeons with long queues. *Anesth Analg* 2009;108:1910–5
41. Xiao Y, Hu P, Hao H, Ho D, Dexter F, Mackenzie CF, Seagull FJ, Dutton R. Algorithm for processing vital sign monitoring data to remotely identify operating room occupancy in real-time. *Anesth Analg* 2005;101:823–9
42. Epstein RH, Dexter F, Piotrowski E. Automated correction of room location errors in anesthesia information management systems. *Anesth Analg* 2008;107:965–71