Prolonged Operative Time to Extubation Is Not a Useful Metric for Comparing the Performance of Individual Anesthesia Providers

Emine Ozgur Bayman, Ph.D., Franklin Dexter, M.D., Ph.D., Michael M. Todd, M.D.

ABSTRACT

Background: One anesthesiologist performance metric is the incidence of "prolonged" (15 min or longer after dressing complete) times to extubation. The authors used several methods to identify the performance outliers and assess whether targeting these outliers for reduction could improve operating room workflow.

Methods: Time to extubation data were retrieved for 27,757 anesthetics and 81 faculty anesthesiologists. Provider-specific incidences of prolonged extubation were assessed by using unadjusted frequentist statistics and a Bayesian model adjusted for prone positioning, American Society of Anesthesiologist's base units, and case duration.

Results: 20.31% of extubations were "prolonged," and 40% of anesthesiologists were identified as outliers using a frequentist approach, that is, incidence greater than upper 95% CI (20.71%). With an adjusted Bayesian model, only one anesthesiologist was deemed an outlier. If an average anesthesiologist performed all extubations, the incidence of prolonged extubations would change negligibly (to 20.67%). If the anesthesiologist with the highest incidence of prolonged extubations was replaced with an average anesthesiologist, the change was also negligible (20.01%). Variability among anesthesiologists in the incidence of prolonged extubations was significantly less than among other providers.

Conclusions: Bayesian methodology with covariate adjustment is better suited to performance monitoring than an unadjusted, nonhierarchical frequentist approach because it is less likely to identify individuals spuriously as outliers. Targeting outliers in an effort to alter operating room activities is unlikely to have an operational impact (although monitoring may serve other purposes). If change is deemed necessary, it must be made by improving the average behavior of everyone and by focusing on anesthesia providers rather than on faculty. **(ANESTHESIOLOGY 2015; XXX:00-00)**

MANY organizations require the assessment of clinical performance metrics although the value of many such metrics remains unknown. Such metrics might be used to evaluate the "quality" of care provided by an individual, assuming that a provider who is identified by such a metric is delivering "lesser quality care" than peers. However, the careless selection and analysis of any given metric may result in individuals being inappropriately labeled as outliers or may result in wasted operational cost.

One proposed metric relevant to both individual "quality" and organizational operations is the incidence of "prolonged" time to endotracheal extubation, which refers to those extubations that occur 15 min or later after the placement of the dressing on the patient.[1–3] They can be measured accurately by both prospective observations in operating rooms (ORs) and retrospectively from anesthesia information management system (AIMS) data.[1,2] There are both clinical and operational reasons for selecting prolonged times to extubation as a potentially meaningful metric:

What We Already Know about This Topic

- Monitoring the incidence of prolonged time to extubation (15 min or longer after dressing applied), a performance metric for anesthesiologists, may be used to identify outliers, provide education, and increase operating room workflow

What This Article Tells Us That Is New

- In a review of over 27,000 anesthetics in a university practice, approximately 20% of extubations were prolonged, with 95% confidence bounds spanning less than 1%
- By a frequentist approach on this small variance data set, 40% of individual anesthesiologists were outliers, whereas with a Bayesian approach only 1% were
- Focusing on changing extubation times only for practitioners who were outliers would have minimal effect on operating room workflow

Corresponding article on page XXX. Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org).

Submitted for publication November 26, 2014. Accepted for publication August 11, 2015. From the Departments of Anesthesia and Biostatistics, University of Iowa, Iowa City, Iowa (E.O.B.); Division of Management Consulting, Department of Anesthesia, University of Iowa, Iowa City, Iowa (F.D.); and Department of Anesthesia, University of Iowa, Iowa City, Iowa (M.M.T.).

<zdoi;10.1097/ALN.0000000000000920>

- Cases with prolonged tracheal extubations are rated by anesthesiologists as having poor recovery from anesthesia.[3]
- Increasing time to extubations increases the chance that at least one member of the OR team (nurses, surgeons, or technicians) will be idle while awaiting extubation ($P < 0.0001$), indicating the slowing of workflow: 21% of teams idle when extubated for less than 5 min, 42% when 5 to less than 10 min, 87% when 10 to less than 15 min, and 100% when 15 min or longer (*i.e.*, prolonged times to extubation).[2]*
- Cases with prolonged times to extubation have substantially (more than 10 min, $P < 0.0001$) longer times from end of surgery to OR exit, even after adjusting for the procedure and positioning.[4]
- Cases with prolonged times to extubation have longer times ($P < 0.0001$) from when the patient exits the OR until the start of surgery of the surgeon's next case in the same OR.[1]
- Because most ($P < 0.0001$) cases with prolonged times to extubation occur during regular workdays and in ORs with greater than 8 h of cases and turnover times, the extra OR time that results can reasonably be treated as an expensive variable cost.[5]
- The incidences of prolonged times to extubations are modifiable (*e.g.*, from meta-analysis of randomized trials, 95% or greater reduction [lower confidence limit] with use of desflurane *vs.* isoflurane).[1,6]
- When surgeons score the importance of anesthesiologists' attributes on a scale from 0, "no importance," to 4, "a factor that would make me switch groups/hospitals," their average score is 3.9 for "patient quick to awaken."[7]

In spite of such information, it is unknown whether the incidences of prolonged times to extubation are a reliable performance metric for individual anesthesiologists and/or for anesthesia providers. We therefore had three goals: (1) extend the processes outlined in our previous work and examine the relative value of frequentist *versus* Bayesian methods for identifying outlier providers for this metric[8]; (2) determine whether efforts at reducing the overall (departmental) incidences of prolonged tracheal extubations would best be achieved by focusing on the subgroups of anesthesiologists and/or anesthesia providers that are performance outliers or rather on the far greater number of providers with incidences close to the overall average; and (3) determine whether most of the heterogeneity was among anesthesiologists or the anesthesia providers that they were supervising.

Materials and Methods

The University of Iowa Institutional Review Board (Iowa City, Iowa) determined that this retrospective quality assurance project concerned primarily clinical activities and did not meet the regulatory definition of human subjects research.

The data gathered were from January 1, 2012 to December 31, 2013 and focused on the 27,788 general anesthetics in which tracheal intubation and extubation were performed in the OR. The details of the structured query language logic to create the analyzed data set are provided in table 1 in the Supplemental Digital Content 1, http://links.lww.com/ALN/B214. Because of changes in the AIMS screens (EPIC; Epic Systems, USA), no earlier data could be used, and the final date was when we started analysis.

The term "anesthesia provider" typically refers to the provider present continually with the patient. In an academic center, these most commonly are residents, nurse anesthetists, fellows, or student nurse anesthetists. For this study, we considered only the anesthesiologist and anesthesia provider present at the time of tracheal extubation, based on staffing information contained in the AIMS. We also limited consideration to residents and nurse anesthetists under the "anesthesia provider" category. Student nurse anesthetists were included only when they were directly supervised by faculty anesthesiologists (not when working one-on-one with a certified registered nurse anesthetist). Similarly, anesthesiology fellows functioned in various roles (sometimes as trainees and sometimes supervising other trainees) and were included only when working under faculty supervision. Only for a small percentage of cases (2.44%) did an anesthesiologist studied personally perform the anesthetic. We included those cases under the "anesthesiologist" category. In order to be included in the performance assessments, the anesthesiologist or the anesthesia provider had to have worked for the department for at least one 6-month period during the 2 studied years. These various restrictions reduced the sample size for performance analyses to 27,757 tracheal extubations for anesthesiologists and 22,086 for anesthesia providers.

Definition of the Prolonged Times to Tracheal Extubation Outcome

The time that the surgical dressing was placed on the patient was defined as the maximum of dressing/cast completion date/time and the procedure end date/time. For 99.72% of cases, both were listed, and the dressing/cast completion time was the later of the two. "Time to extubation" in this data set was defined as the time from dressing complete to the recorded time of endotracheal tube removal.

Selection of Covariates for Bayesian Analyses

In the current study, Bayesian models for anesthesiologists and anesthesia providers were fit separately under two different conditions: (1) no adjustment and (2) model adjusting for patient covariates based on the classification tree analyses. Adjustments

* Our previous observer study (available at: http://FDshort.com/Masursky2012) included video (Supplemental Digital Content, http://links.lww.com/AA/A396) showing animation of a typical observation period, highlighting that 15 min is a *very* long time in terms of operating room activity.

Table 1. Descriptive Statistics for Variables Used in the Model

Variables	n	Statistics
Time from OR entrance until the dressing has been placed (min)	27,788	
Mean ± SD		181.19 ± 110.11
Median (Q_{25}, Q_{75})		158.00 (103.00, 233.00)
American Society of Anesthesiologists' Base Units of the primary surgical procedure performed	25,944	
Mean ± SD		6.39 ± 2.71
Median (Q_{25}, Q_{75})		6.00 (5.00, 7.00)
	% missing	(%) N
Patient's last position	0%	
Prone		5.79 (1,610)
Not prone		94.21 (26,178)
American Society of Anesthesiologists' Base Units of the primary surgical procedure performed	0%	
< 11		92.30 (25,647)
≥ 11		7.70 (2,141)

American Society of Anesthesiologists' Base Units of the primary surgical procedure performed of 11 includes nearly all intracranial neurological procedures.

OR = operating room; Q_{25} = 25th percentile; Q_{75} = 75th percentile.

for the models were performed as follows. First, a large data set consisting of *all* the preoperative characteristics in the AIMS for all the patients with tracheal intubation and extubation when the patient was in an OR (128 variables) was collected. Second, classification tree analyses were performed by using SAS Enterprise Miner software 7.1 (SAS Institute Inc., USA). Models were compared based on the mean squared error.

Classification/decision tree analysis consists of a hierarchy of branches.[9] Each branch was divided to up to two branches. The same variable was not used in more than one branch.

In a previous study, we found that two principal predictors of prolonged times to extubation were (1) surgery performed in the prone position and (2) whether the time from OR entrance until the dressing (a measure of case duration) has been placed was 4 h or longer.[4] For the current study, all 128 preoperative and intraoperatively available covariates were used in the SAS Enterprise Miner (table 1, and tables 2 and 3 in the Supplemental Digital Content 1, http://links.lww.com/ALN/B214). The predictors identified were the same two plus (3) whether the numbers of American Society of Anesthesiologists' Relative Value Guide base units were 11 units or greater. The 11-unit cases were principally that of intracranial surgery (fig. 1).

The Bayesian method uses logistic regression models. Prone position (prone *vs.* not prone) and the numbers of American Society of Anesthesiologists' Relative Value Guide base unit of the procedure being greater than or equal to 11 units (11 units or greater *vs.* less than 11 units) are binary variables (yes/no). Time from OR entrance until the dressing has been placed is the only continuous variable in the model. Even if the time from OR entrance until the dressing has been placed was divided to two branches in the decision tree, in the actual analysis, this variable was used as a continuous variable. Box–Cox transformation was used to determine the best transformation for the time from OR entrance until the dressing has been placed to satisfy the assumption of a linear relation between the transformed variable and the incidence of prolonged time to extubation on the logit scale. Two times the square-root transformation provided the best result for this variable (*i.e.*, closest to linear relation with the incidence of prolonged time to extubation on the logit scale).

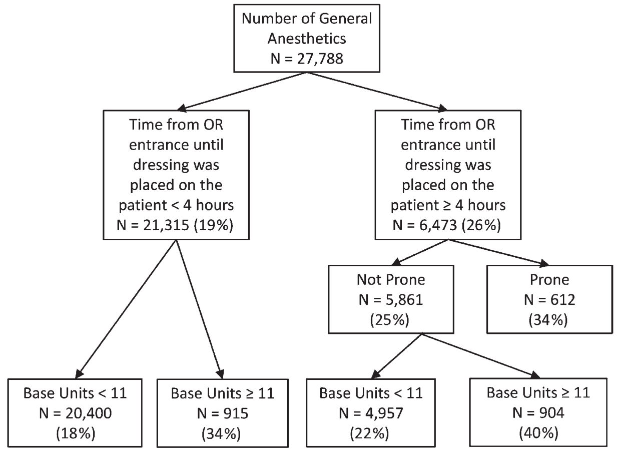
Number of General Anesthetics N = 27,788
Time from OR entrance until dressing was placed on the patient < 4 hours N = 21,315 (19%)
Time from OR entrance until dressing was placed on the patient ≥ 4 hours N = 6,473 (26%)
Not Prone N = 5,861 (25%)
Prone N = 612 (34%)
Base Units < 11 N = 20,400 (18%)
Base Units ≥ 11 N = 915 (34%)
Base Units < 11 N = 4,957 (22%)
Base Units ≥ 11 N = 904 (40%)

Fig. 1. SAS Miner decision tree (SAS Institute Inc., USA) for the incidence of prolonged times from the end of surgery (dressing on patient) until tracheal extubation. The time from dressing on the patient to removal of endotracheal tube was considered prolonged if 15 min or longer. SAS Miner decision tree determined the cutoff from entrance into the operating room (OR) until the final dressing was placed to be 242 min. For ease of interpretation, we rounded this to 240 min (*i.e.*, to 4 h). This cutoff value matched the results from table 1 of the article by Dexter and Epstein,[4] obtained using data from a different hospital. Although the time from OR entrance until the dressing has been placed was divided to two branches in the decision tree, this variable was used as a continuous variable in the actual analyses. The American Society of Anesthesiologists' Base Units of the primary surgical procedure performed are not ratio levels of measurement, but ranked with respect to extubation time. The cutpoint of 11 base units achieved the least mean square error when it was chosen as the third variable in the decision tree. Intracranial neurological procedures have 11 base units.

In the logistic regression model, all main-effects terms as well as all two- and three-way interaction (prone × American Society of Anesthesiologists base unit × case duration) terms were tested. The three-way interaction model was significant, and two-way interactions were not significant. The *c*-statistics for this full model was 0.557. When the three-way interaction term was not in the model, two of the two-way interaction terms stayed nonsignificant. After removing the nonsignificant two-way interaction terms stepwise from the model,

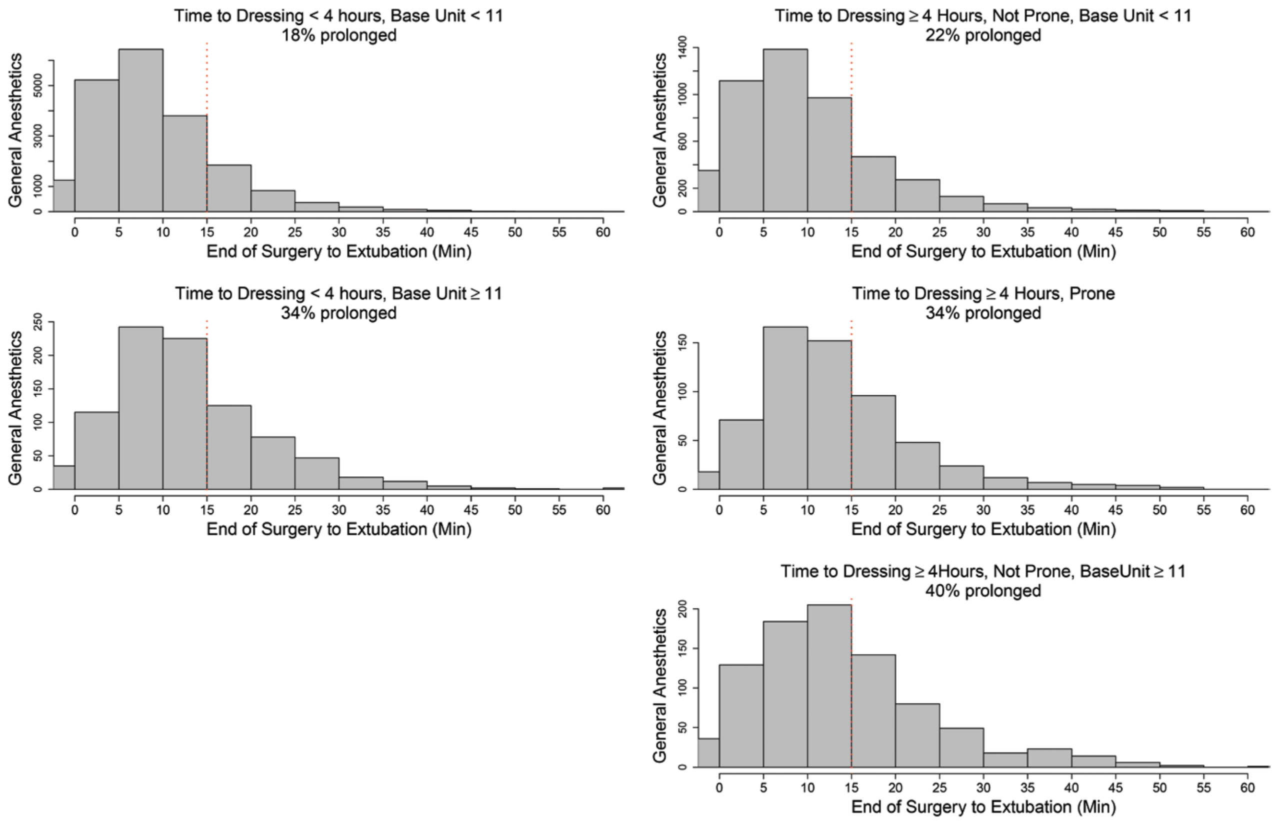

Fig. 2. Histogram of time to extubation for the five entries of figure 1. The incidences for "prolonged time to extubations," when the time to extubation was 15 min or longer, are also presented for each panel. Time to dressing: time from operating room entrance until dressing was placed on the patient.

all the main effects (the prone position, the time from OR entrance until the dressing has been placed, and base units were 11 or greater) became significant in addition to the prone and base unit interaction term. The *c*-statistics of this model was identical to the *c*-statistics of the full model (0.557). The model with main-effects and one interaction term was used (the prone position, the time from OR entrance until the dressing has been placed, and base units were 11 or greater and prone × base units), instead of the full model, as it was more parsimonious, and easier to interpret. This reduced model was also consistent with the decision tree produced by the SAS Enterprise Miner (fig. 1). Histograms showing the distribution of time to extubation for each of the five branches in figure 1 are presented in figure 2.

All 125 of the covariates that were not used in the model were tested individually against the reduced model (main-effects model of prone position, the numbers of American Society of Anesthesiologists' Relative Value Guide base units of greater than or equal to 11, and the OR in dress time and the interaction of prone and base unit) to test whether any of the other covariates make a meaningful increase on the area under the curve. The individual improvements on the area under the curve, after the inclusion of other variables, in addition to the existing model, were all less than 3.1% absolute increment. For example, although Dexter and Epstein[4] did not find age to be predictive for prolonged time to extubation, it is a common covariate, and thus we investigated it in detail; its resulting absolute increase in the area under the curve was only 0.23%. Therefore, the preceding logistic regression model was used in the Bayesian hierarchical generalized linear model.

Frequentist Outlier Detection Methods

Methods for modeling anesthesiologists are explained in this section. The same methods are applied below for modeling anesthesia providers.

The first objective of the current article was to identify those anesthesiologists with a significantly greater incidence of prolonged tracheal extubations than the other anesthesiologists. This was needed to learn what percentage of prolonged tracheal extubations was attributable to outlier anesthesiologists. To do this, "outlier" needed to be defined.

There are multiple frequentist criteria for defining an outlier. For example, Ehrenfeld *et al.*[10] defined the "worst" 5% of all anesthesiologists (on any metric) as being outliers, regardless of their specific performance. An alternative is to examine the overall incidence of a given event (*e.g.*, prolonged time to extubation), calculate confidence bounds around that incidence, and define any provider whose individual incidence is beyond the 95% upper bound to be an outlier. Typically, such frequentist metrics are unadjusted. If an individual's incidence is greater than the boundary, by any margin, they are considered an outlier, regardless of case numbers, case types, or other factors. For this article, we examined the upper one-sided 95% CI of the overall incidence of prolonged tracheal extubations and treated this CI as a threshold for frequentist outlier although we recognize that other definitions might be used.

Bayesian Outlier Detection Methods

The method developed by Chaloner and Brant[11] and applied by Bayman *et al.*[12] was used to identify outliers. Bayman *et al.*[12] introduced a Bayesian approach to detect the outliers among centers in multicenter clinical trials. In the current study, their method was applied to detect anesthesiologists with incidences of prolonged times to extubation that were different from other anesthesiologists, while taking into account patient and procedure characteristics. Adjusted and unadjusted models were fit separately for each of anesthesiologists and anesthesia providers. The same covariates were used for model 1 including anesthesiologists and model 2 including anesthesia providers.

It was assumed that each tracheal extubation was performed by a single anesthesiologist, either an anesthesiologist for model 1 or nurse anesthetists or residents for model 2. In mathematical terms, anesthetics were nested within anesthesiologists. Bayesian hierarchical generalized linear models were fit for the prolonged time to extubation outcome. Details of the model are given in appendix A.1.[13,14]

In Bayesian analyses, unknown parameters are random variables and therefore prior probability distributions should be defined. The Bayesian model combines the prior distribution with data and produces a posterior distribution. Inferences are made from the posterior distribution. Bayesian credible intervals, analogous to frequentist CIs, were constructed based both on the prior information and the observed data.

Two different prior probabilities were examined for an anesthesiologist having an incidence of prolonged time to extubation that was so different from others that he or she appeared to be an outlier: (1) the prior probability of each anesthesiologist having an outlier incidence of prolonged tracheal extubations was set to 5% (appendix A.2).[11] (2) The prior probability of no anesthesiologist in the department being an outlier over 2 yr was set to 95% (appendix A.3).[11] With 81 anesthesiologists in the department during the study period, the prior probability of each anesthesiologist being an outlier was 0.06% (appendix A.3).[11]

Prior distributions used for the overall mean, and the coefficients for the fixed-effect terms such as patient's prone position, the time from OR entrance until dressing was placed on the patient, and whether the case's number of American Society of Anesthesiologists' Relative Value Guide base unit was 11 units or greater were assumed to be normally distributed, as usual for these types of analyses, and were weakly informative (have very large SDs). Random prior distributions were defined for each anesthesiologist.

Posterior probabilities of being an outlier were calculated for each anesthesiologist, and the strength of evidence was quantified by the Bayes factor.[14] Bayes factor is the ratio of the posterior odds in favor of the null to the prior odds of the null.[15] The most common interpretation of Bayes factor classifies evidence against the null hypothesis as "strong," "very strong," and "decisive" when the Bayes factors were less than 10^{-1}, $10^{-1.5}$, and 10^{-2}, respectively, according to Jeffreys scale.[14] Kass and Raftery[16] recommend a more conservative interpretation where Bayes factors less than 0.33, 0.05, and 0.0067 are classified as "positive," "strong," and "very strong" evidence against the null hypothesis. With both scales, Bayes factor greater than 1 provides evidence *for* the null hypothesis. The overall posterior probability for at least one of the anesthesiologists being an outlier was also calculated. An anesthesiologist with a Bayes factor less than 0.1, which indicates "strong" evidence according to the Jeffreys scale,[14] was identified as an outlier.

Bayes factors for those anesthesiologists (or anesthesia providers) with a significantly greater or lesser incidence of prolonged times to extubations than the other anesthesiologists are provided in the figure legends. In addition, the strength of evidence was provided among outlier anesthesiologists with "strong," "very strong," and "decisive" evidence. It should be noted that, an anesthesiologist who is an outlier with a "decisive" evidence is also an outlier with a "very strong" and "strong" evidence. The direction of the outlier anesthesiologist, with significantly greater or significantly less incidence than the rest of the anesthesiologists, was determined by the sign of the random anesthesiologist effect (δ_k). A negative δ_k indicates a greater incidence of prolonged times for tracheal extubation for the *k*th anesthesiologist. Anesthesiologists (or anesthesia providers) with both significantly greater and significantly less incidence of prolonged times to extubations are reported. Analyses were repeated by using different prior distributions as sensitivity analyses (appendix A.4)[14,17] (for statistical details and explanations of the WinBUGS model, see the appendix A.5).[18]

The log odds of not having a prolonged time to extubation for the *i*th endotracheal extubation for anesthesiologist *k* (θ_{ik}) for the adjusted Bayesian model when each anesthesiologist's prior probability of being an outlier was set to 5% with posterior means substituted as estimates for the coefficients is as follows:

$$\theta_{ik} = 1.81 - 0.36\text{Prone} - 0.75\text{BaseUnit} - 0.01\text{ORDrTime} \\ + 0.20\text{Prone} \times \text{BaseUnit} + \delta_k .$$

See appendix A.1 for the explanations of the model terms. There are 81 different values of the random anesthesiologist effect term, δ_k. They range from −0.9555 to 0.4990 on the logit probability scale.

For the second goal to determine whether targeting outliers might have a meaningful impact on OR workflow, further calculations were performed as follows. For calculations to represent a typical patient, probabilities of prolonged tracheal extubations were calculated based on all 27,757 anesthetics. To calculate the incidence if all patients were cared for by the average anesthesiologist, the pooled estimate of the random anesthesiologist effects was used (appendix A.6). To calculate the incidence of prolonged tracheal extubations if the anesthesiologist with the largest adjusted incidence of prolonged times to extubations cared for every patient, this provider's random provider effect ($\text{delta}_{35} = -0.9555$) was used for all 27,757 anesthetics.

Basic data analyses were performed by using SAS software 9.3, and classification tree analyses were performed by using SAS Enterprise Miner software 7.1 (SAS Institute Inc.). Plots were created using SigmaPlot version 12.5 (Systat Software, USA) and R version 3.0.0 (The R Foundation, Austria).[19] Bayesian analyses were performed by using WinBUGS 1.4.3 software.[18] WinBUGS uses Markov chain Monte Carlo methods. To represent the extreme regions of the parameter space, three parallel chains of equal lengths with disperse initial values were used in WinBUGS analyses. Convergence was judged by Brooks, Gelman, Rubin diagnostics plots,[20] density and history plots, and autocorrelations.

Bayesian results were based on 5,000 iterations after a burn-in period of 5,000 iterations in each chain (Supplemental Digital Content 2, http://links.lww.com/ALN/B215).

To test whether (1) anesthesiologists or (2) anesthesia providers have greater variability in the incidence of prolonged extubation, the Convergence Diagnostic and Output Analysis option of WinBUGS program was used. For both adjusted models from anesthesiologists and anesthesia providers, between-anesthesiologist or between-anesthesia provider SDs were monitored for all 15,000 replications. The two groups each with 15,000 replications were compared by the Wilcoxon rank sum test. The Wilcoxon–Mann–Whitney odds (WMWodds) was used as a summary measure for the Wilcoxon rank sum test.[21,22]

Reporting of Bayes Used in Clinical Studies guidelines was followed to report Bayesian analyses in this study.[23]

Results

The overall incidence of prolonged times to extubation among 27,757 cases with anesthesiologists was 20.31% (the upper 95% CI, 20.71%) (table 2). The incidence based on the 22,086 cases with an anesthesia provider present (*i.e.*, anesthesiologist not personally performing the case) was 19.75% (the upper 95% CI, 20.20%) (table 3).

Summary results and unadjusted (raw) incidences of prolonged tracheal extubations are given in table 2 and figure 3 for anesthesiologists and in table 3 and figure 4 for anesthesia providers. For example, 81 anesthesiologists performed 27,757 extubations, and the number of extubations per anesthesiologist ranged from 13 to 1,079.

Comparisons Based on Frequentist Methods

When the frequentist method (described in the Frequentist Outlier Detection Methods section) was used, those anesthesiologists with an incidence of prolonged extubations greater than 20.71% (the upper 95% CI) would be classified as outliers. However, this corresponded to 40% (32 of 81) of the anesthesiologists in our department (table 2). Similarly, incidences

Table 2. Summary Results for Anesthesiologists for the Incidences of Prolonged Times to Tracheal Extubations between January 1, 2012 and December 31, 2013

	Each Anesthesiologist's Prior Probability of Being Outlier = 5%	The Departmental Prior Probability of Any Outlier Anesthesiologist = 5%
Number of anesthetics evaluated	27,757	
Number of evaluated anesthesiologists supervising at least one anesthetics	81	
Number of anesthetics per anesthesiologists	13 to 1,079	
The incidence of evaluated anesthetics with noncompliance	20.31%	
Anesthesiologists identified as performance outliers		
Frequentist	n = 32 of 81	
	Significantly greater incidence	
Bayesian unadjusted	n = 2 of 81	n = 1 of 81
Significantly greater incidence	(31.33%, 42.50%) one strong (no. 4), one decisive (no. 35)	(42.50%) one decisive (no. 35)
Significantly lesser incidence	0 of 81	0 of 81
Bayesian adjusted	n = 2 of 81	n = 1 of 81
Significantly greater incidence	(30.27%, 42.50%) one strong (no. 6), one decisive (no. 35)	(42.50%) one decisive (no. 35)
Significantly lesser incidence	0 of 81	0 of 81

The between-anesthesiologist variance (the variance of δ_k's) for the adjusted Bayesian model with the individual prior probability is 0.314^2 on the logit probability scale. The random anesthesiologist effects change between −0.956 and 0.499. For the anesthesiologist with a significantly greater adjusted incidence of prolonged time to extubation than the other anesthesiologists, δ_k is −0.956 on the logit probability scale.

Table 3. Summary Results for Anesthesia Providers (Certified Registered Nurse Anesthetists/Residents) for the Incidences of Prolonged Times to Tracheal Extubations between January 1, 2012 and December 31, 2013

	Each Anesthesia Provider's Prior Probability of Being Outlier = 5%	The Departmental Prior Probability of Any Outlier Anesthesia Provider = 5%
Number of anesthetics evaluated	22,086	
Number of evaluated anesthesia providers supervising at least one anesthetics	116	
Number of anesthetics per anesthesia providers	20 to 495	
The incidence of evaluated anesthetics with noncompliance	19.75%	
Anesthesia providers identified as performance outliers		
Frequentist	n = 63 of 116	
	Significantly greater incidence	
Bayesian unadjusted	n = 0 of 116	n = 0 of 116
Significantly greater incidence	(Not applicable)	(Not applicable)
Significantly lesser incidence	5 of 116: 4 decisive (nos. 15, 18, 23, and 24) and 1 strong (no. 4)	4 of 116: 1 decisive (no. 24), 3 strong (nos. 15, 18, and 23)
Bayesian adjusted	n = 1 of 116	n = 0 of 116
Significantly greater incidence	(41.35%) 1 strong (no. 70)	(Not applicable)
Significantly lesser incidence	5 of 116: 2 decisive (nos. 23 and 24), 2 very strong (nos. 15 and 18), and 1 strong (no. 4)	3 of 116: 1 decisive (no. 24), 2 strong (nos. 15 and 23)

For the adjusted Bayesian model with the individual prior probability, random anesthesia provider effects change between −1.041 and 1.653 on the logit probability scale. The between-anesthesia provider variance for this model is 0.558.[2]

of prolonged times to endotracheal extubation were significantly greater than 20.20% (the upper 95% CI) for more than half (54%, 63 of 116) of the anesthesia providers (table 3). These "outliers" are presented with blue hexagons in figures 3 and 4. Because such numbers stretch the meaning of the word "outliers," we do not rely on this approach any further.

Comparison among Anesthesiologists Based on the Bayesian Approach

When the Bayesian model was used without adjusting for any patient and surgical covariate, and each provider's prior probability of being outlier was set to 5%, two anesthesiologists (nos. 4 and 35) were identified as having a significantly greater incidence of prolonged times to extubation than the other anesthesiologists (first column of table 2) (nos. 4 and 35 refer to the anesthesiologists with the 4th and 35th largest numbers of tracheal extubations over the 2 yr). Observed incidences of prolonged times to extubations for these anesthesiologists were 31.33 and 42.50%. The posterior probabilities of these anesthesiologists being outliers were 42 and 99%.

When the Bayesian model was used after adjusting for prone positioning, the numbers of American Society of Anesthesiologists' Relative Value Guide base unit, and the time from OR entrance until dressing was placed on the patient, two anesthesiologists (nos. 6 and 35) were identified as having significantly greater adjusted incidence of prolonged times to extubation than the other anesthesiologists (table 2 and fig. 3). The posterior probabilities for these two anesthesiologists being outlier were increased from 5% to 47% and 99%, respectively.

There was no anesthesiologist with a significantly lesser incidence of prolonged tracheal extubation times than other anesthesiologists, both by unadjusted and adjusted Bayesian models.

To quantify the variability among anesthesiologists, we statistically treated each of the 81 anesthesiologists as caring for the same representative patient. This representative patient was defined as one that underwent a procedure in any position other than prone, with 10 or fewer American Society of Anesthesiologists' Relative Value Guide base units, and a case duration of 158 min or less. Random anesthesiologist effects were used from the adjusted model, with each provider's prior probability of being outlier was set to 5%. Figure 5 shows the plot of rank of anesthesiologists and the associated 95% credible intervals (vertical axis) by the incidence of prolonged tracheal extubations (horizontal axis) for this representative patient. Those anesthesiologists with greater incidences of prolonged times to extubations were ranked lower and can be found on the left side of the figure.

The zoomed version of figure 5 for those anesthesiologists with 20% or greater incidence of prolonged times to extubations is provided in figure 6. The one anesthesiologist with a significantly greater adjusted incidence of prolonged times to extubation for both unadjusted and adjusted model (no. 35) is ranked 1 and represented with a red square on this figure. The associated 95% credible interval of the rank is narrow. The anesthesiologist who was detected as an outlier according to the unadjusted model, but not adjusted model, is presented with a solid green triangle. The anesthesiologist detected as an outlier according to the adjusted model, but not the unadjusted model, is presented with a

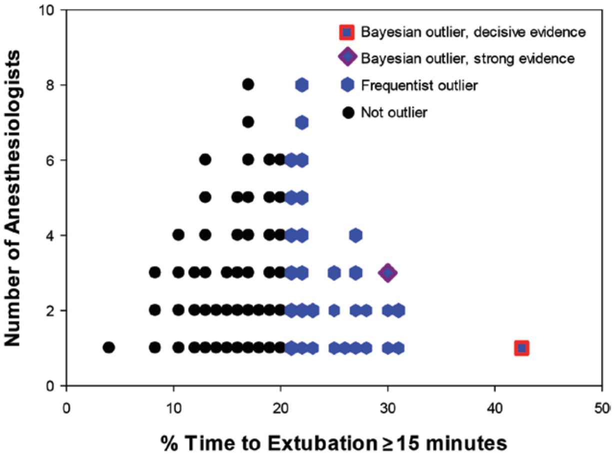

Fig. 3. Dotplot for prolonged time to extubation for cases ending January 1, 2012 through (and including) December 31, 2013 and classified by anesthesiologist. Each anesthesiologist's prior probability of being an outlier was set equal to 5%. Bayes factors (BF) for those anesthesiologists with a significantly greater or less incidence of prolonged times to extubations than the other anesthesiologists are provided. For example, BF_1 represents the BF of the first anesthesiologist (the anesthesiologist with the greatest number of tracheal extubations in 2 yr). These results using the adjusted model show two outlier anesthesiologists. One anesthesiologist was an outlier with strong evidence (BF_6 = 0.06). The other anesthesiologist was an outlier with decisive evidence (BF_{35} = 0.0001). That second anesthesiologist who was detected as an outlier with decisive evidence was, by definition, also an outlier with "very strong" and "strong" evidence. Note that, these two anesthesiologists were also detected as frequentist outliers, which is why their *symbols* also include some *blue*. The posterior probability for the department having at least one outlier anesthesiologist was 99.9%. The *far left side* of the figure shows data from an anesthesiologist with an adjusted incidence of prolonged time to extubation of 4%. This anesthesiologist was not detected as an outlier (*i.e.*, did not have a significantly less incidence of prolonged times to extubation than the other anesthesiologists). This anesthesiologist's incidence was still within the normal variability expected from a normal distribution among the anesthesiologists.

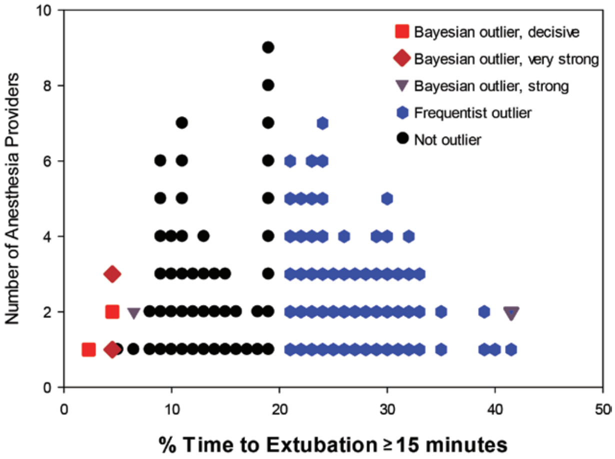

Fig. 4. Dotplot for prolonged time to extubation for cases ending January 1, 2012 through (and including) December 31, 2013, and classified by anesthesia provider (resident or nurse anesthetist) according to the adjusted Bayesian model. Each anesthesia provider's prior probability of being an outlier was set equal to 5%. One provider was detected as having significantly greater adjusted incidence of prolonged time to extubation than the other providers with strong evidence (Bayes factor $[BF]_{70}$ = 0.08). There were five anesthesia providers with a significantly less incidence of tracheal extubations than other anesthesia providers; two with decisive evidence (BF_{23} = 0.008, BF_{24} = 0.0009), two with very strong evidence (BF_{15} = 0.010, BF_{18} = 0.016), and one with strong evidence (BF_4 = 0.06). Incidences of prolonged times to extubations for these providers were between 2.32 and 6.80%.

blue square. Importantly, the credible intervals in the figure are *not* adjusted for multiple comparisons[24] (*i.e.*, actual credible intervals would be even wider than the already wide, displayed intervals). This highlights the strength of evidence that there was one outlier, anesthesiologist no. 35; none of the other 80 anesthesiologists differed significantly from one another in their incidences of prolonged times to tracheal extubation.

Figure 7 provides another view of the information in figures 5 and 6. The figure 7 also shows the substantial overlap of the posterior incidences of prolonged tracheal extubations among anesthesiologists. In figure 7, the posterior distributions of the ranks of five selected anesthesiologists are displayed, those being the anesthesiologists with representative patient incidences of prolonged times to extubation of 35, 25, 20, 15, and 10%, respectively. For example, figure 7A shows the simulated relative ranks for the anesthesiologist with the incidence of prolonged times to extubation of 35% for the representative patient. The posterior rank distribution of that anesthesiologist no. 35 is centered on the rank of 1 with a small SD and in only some replications having a rank of 2. In other words, for almost all replications, this anesthesiologist's rank was 1, indicating the greatest incidence of prolonged times to extubation. However, for all other anesthesiologists, there were substantial overlaps of credible intervals of ranks. The latter is the important finding because it applies to 80 of 81 anesthesiologists.

Based on these results, we concluded that the anesthesiologist no. 35 was truly an outlier.

For the analyses for the second goal, using each patient's three variables (prone position, numbers of American Society of Anesthesiologists' Relative Value Guide base units, and time from OR entrance until the dressing has been placed), the probability of prolonged tracheal extubations was recalculated for every case under five different scenarios: (1) no change; (2) if all anesthesiologists had the overall performance of the average anesthesiologist, (3) all anesthesiologists had the performance of the sole outlier anesthesiologist (*i.e.*, anesthesiologist no. 35 with the greatest adjusted incidence of prolonged times to extubations), (4) the anesthesiologist with the greatest adjusted incidence of prolonged times to extubations (anesthesiologist no. 35) was replaced

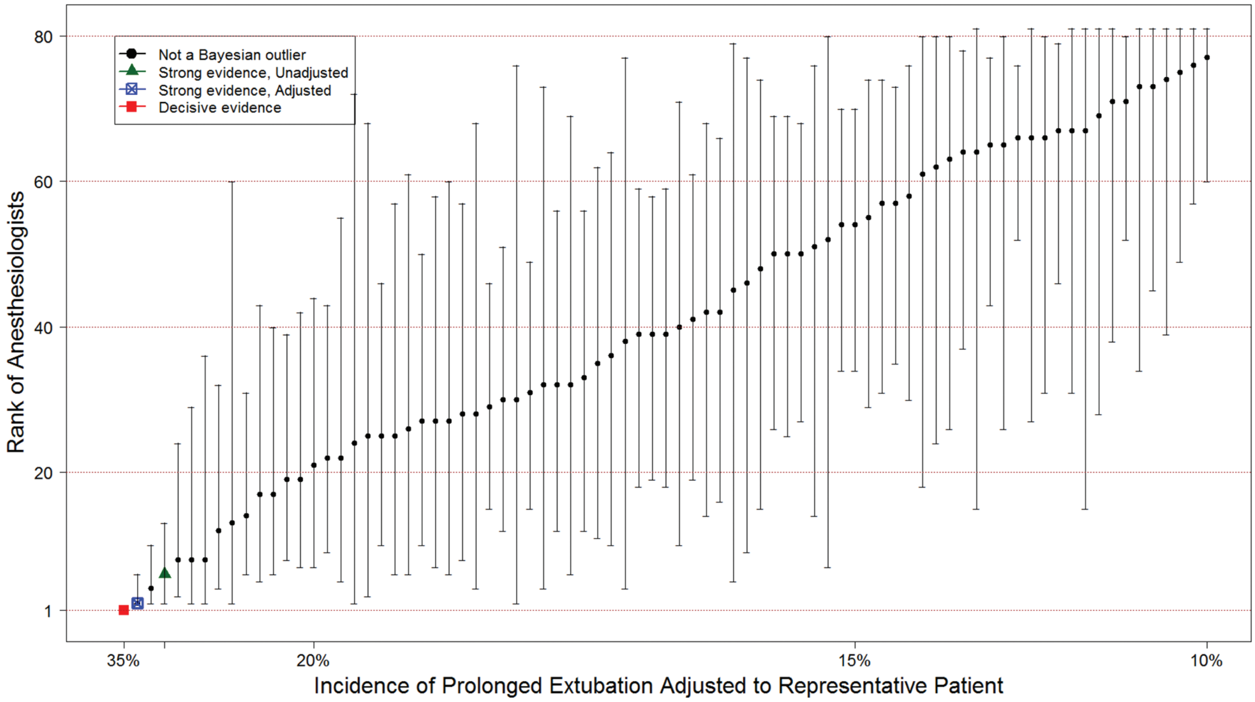

Fig. 5. Using the median of 158 min based on all 27,757 anesthetics. Rank of the incidences of prolonged times to extubation among anesthesiologists. The ranks are provided on the vertical axis of the figure. Those anesthesiologists with greater incidences of prolonged times to extubation have been assigned lesser ranks (*i.e.*, the anesthesiologist with rank 1 had the greatest adjusted incidence of prolonged times to tracheal extubation). That anesthesiologist has decisive evidence of being an outlier (*red square*). Anesthesiologists who tend to have lesser incidences of prolonged times to tracheal extubation compared with the other anesthesiologists are cumulated toward the right hand side, toward the rank of 80. For example, for the anesthesiologist on the far right hand side, the 95% credible interval ranges between 60 and 80. That result means that the rank of this particular anesthesiologist was between 60 and 80 for 95% of the 15,000 replications. In other words, that anesthesiologist's incidence of prolonged time to tracheal extubation was less than most other anesthesiologists. The differences between the observed and adjusted incidences of prolonged times to tracheal extubation were small (mean = 1.3%, SD = 3%) and were the largest for those anesthesiologists on the left tail area for all 81 anesthesiologists. For those 10 anesthesiologists at each extreme of the distribution (five on each side), percentage of the times the anesthesiologist personally performed the extubation (*i.e.*, there was no anesthesia provider present) was similar (medians are less than 5%).

by the average anesthesiologist's effect, and (5) the performances of the 32 frequentist outlier anesthesiologists were replaced by that of the average anesthesiologist.

If each of 81 anesthesiologists own random-effect term was used, the average probability of prolonged times to extubation became 20.24%. If an average anesthesiologist performed all 27,757 extubations, the average incidence would increase only 0.43%, a clinically and managerially unimportant change (standard errors [SEs] for this and the next four results were all less than 0.2%). In other words, as suggested in figures 5 and 7, most anesthesiologists performed very similar to the average. In contrast, if all anesthesiologists had the same incidence as the outlier anesthesiologist, with the greatest incidence, the incidence of prolonged endotracheal extubations would be 37.63% (*i.e.*, almost doubled). This is substantial, and the implication is that the one outlier anesthesiologist had substantially longer extubation times than the other anesthesiologists. Still, there was just one individual outlier anesthesiologist. When that anesthesiologist's random effect (*i.e.*, personal adjusted incidence) was replaced with the weighted average of all anesthesiologists, the overall incidence became 20.01%, just 0.23% less than when using anesthesiologists' own random-effect term (20.24%). This 0.23% reduction is the principal (practical) question of importance because monitoring for outlier anesthesiologists would be useful only if this difference had been substantial. If the 32 frequentist outlier anesthesiologists' performances were replaced by the weighted average of all anesthesiologists, the average probability of prolonged times to extubation would be 18.92%, only a 1.32% reduction from 20.24%. This result implies that we should try other things to improve this outcome.

The analysis by anesthesiologist included n = 27,757 cases, but that of residents and nurse anesthetists included n = 22,086 cases because of cases with anesthesiologist only, student nurse anesthetists, or fellows. The adjusted model was rerun using just the 22,086 cases, and with the prior probability of each anesthesiologist being an outlier equal to 5%. The results were indistinguishable, with the SD for the random anesthesiologist effect in the logit scale being 0.317 (SE = 0.030) instead of 0.314 (SE = 0.029).

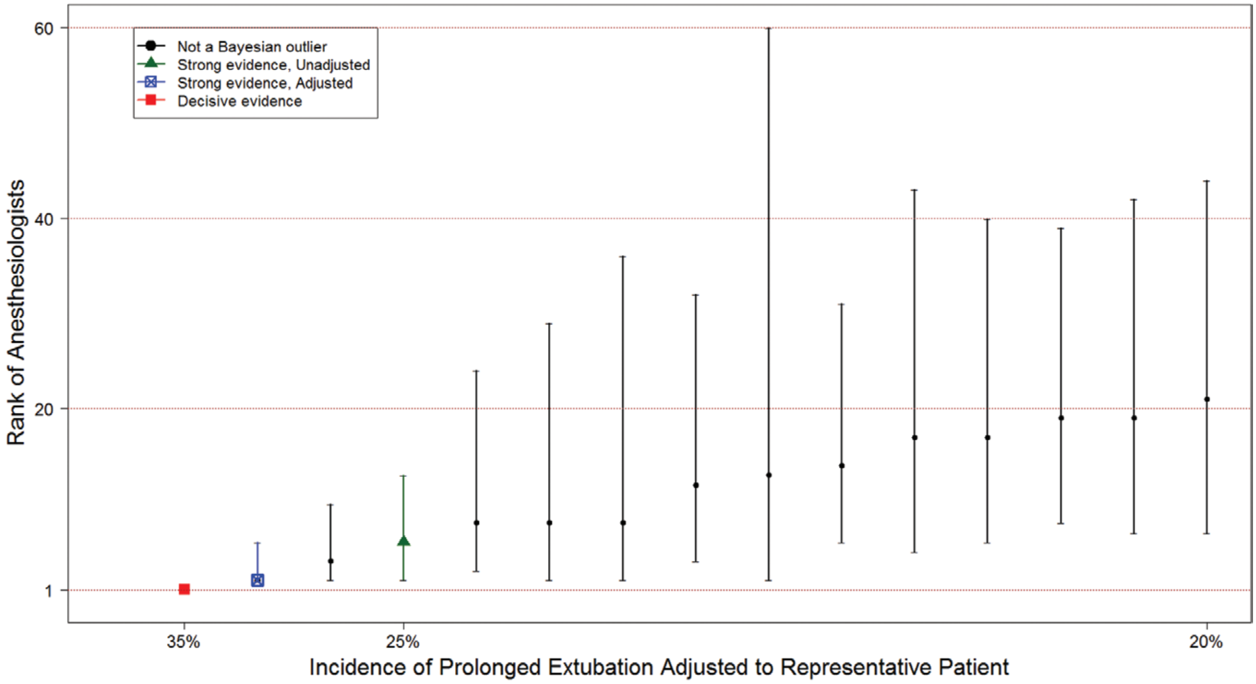

Fig. 6. Zoomed version of the first 15 anesthesiologists with 20% or higher incidence of prolonged times to extubations adjusted to a representative patient.

Comparison among Anesthesia Providers Based on the Bayesian Approach

The same methods were applied to the extubation times of the 116 anesthesia providers, resident physicians, and nurse anesthetists. The overall incidence of prolonged tracheal extubation times for the 22,086 tracheal intubation and extubations performed by this group was 19.75%. There was not even one anesthesia provider with a significantly greater incidence of prolonged times to extubation than the other anesthesia providers, according to the unadjusted Bayesian model when each provider's prior probability of being outlier was set to 5% (first column of table 3). When the adjusted Bayesian model was used, one provider was detected as having significantly greater adjusted incidence of prolonged times to extubation than the other providers (no. 70). The posterior probability for this anesthesia provider being an outlier was 39.5%.

Variability among Anesthesiologists *versus* among Anesthesia Providers

Anesthesiologists (and anesthesia providers) were assumed to have a random normal distribution with mean 0 and SD of σ, where σ represents the within-group (within anesthesiologists or within anesthesia providers) variability of the logit of the probability of prolonged time to extubation. When between-anesthesiologist SDs of the random effects identifying the provider were compared by using adjusted models of (1) anesthesiologists $\left(\sigma_{\text{Anesthesiologist}} = 0.314\right)$ *versus* (2) anesthesia providers $\left(\sigma_{\text{Anesthesia provider}} = 0.558\right)$, a greater variability was observed among the anesthesia providers ($P < 0.00001$ for both two independent-samples t test and Wilcoxon rank sum test, see fig. 1 in Supplemental Digital Content 3, http://links.lww.com/ALN/B216, for a box plot). The odds of the SD being less among anesthesiologists than anesthesia providers was at least 1,000 (WMWodds lower 95% CI). When the same calculations were performed using a noninformative prior distribution for the SD of the random anesthesiologist effect, or setting the individual prior probability of being outlier to 1% or 10% for each anesthesiologist and anesthesia provider, the WMWodds was still at least 1,000 (lower 95% CI). This shows that the greater variability in incidences among anesthesia providers than among anesthesiologists is robust to the selected prior distributions. The greater variability among anesthesia providers *versus* anesthesiologists can be seen graphically. For example, in figures 3 and 4, the incidences of prolonged times to endotracheal extubation were between 10 and 30% among 94% of anesthesiologists *versus* among 72% of anesthesia providers. Figure 8 is a quantile-quantile plot comparing the distribution of the incidences of prolonged times to endotracheal extubation among anesthesiologists *versus* among anesthesia providers (see Discussion).

Discussion

Recent years have seen an increase in the development and application of provider and group-focused metrics. Such metrics may be intended to identify individual providers whose performance differs from either their peers or from various benchmarks. They may also be used to guide administrative efforts to alter group practices. Although all such metrics are *intended* to improve patient care, it is not always apparent

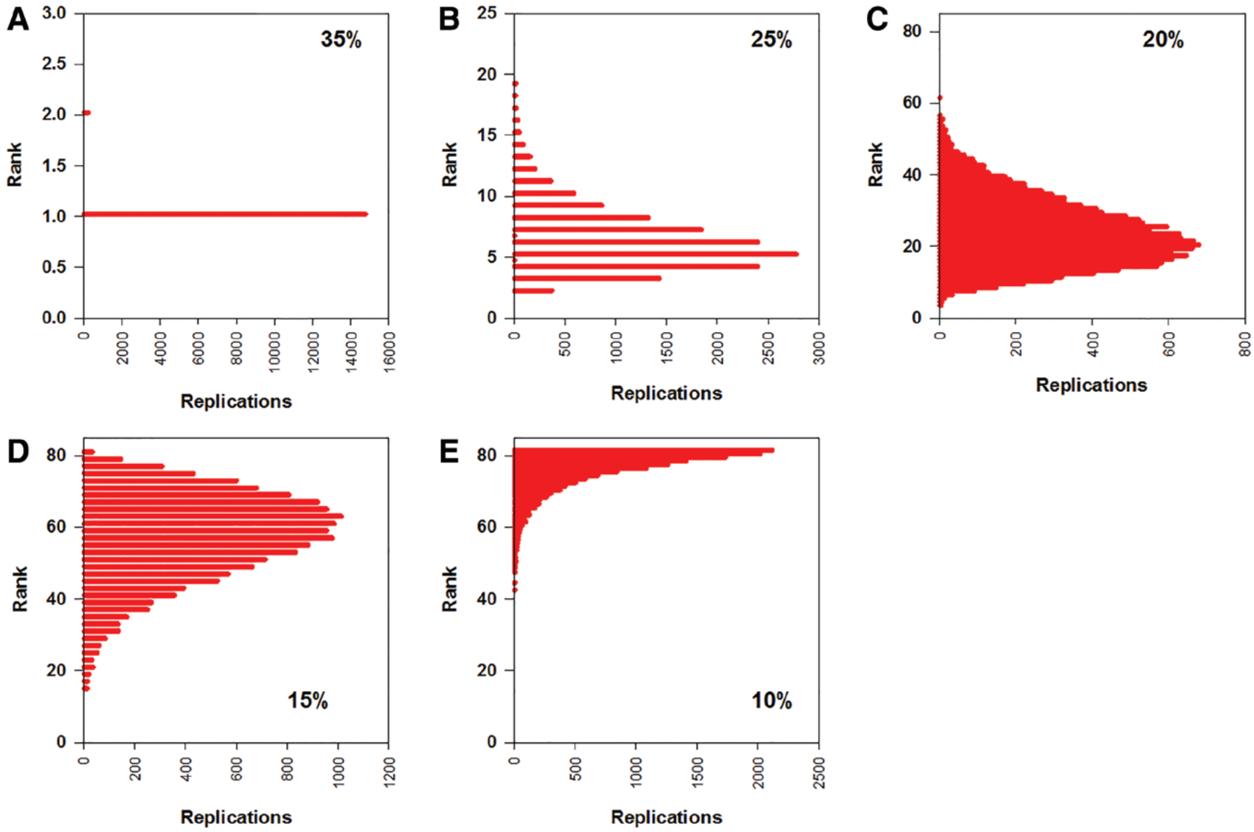

Fig. 7. Posterior distribution of the rank among five anesthesiologists with selected adjusted incidences of prolonged times to tracheal extubation. The posterior distributions of the ranks of five selected anesthesiologists are displayed in *A* to *E*; those being the anesthesiologists with representative patient incidences of prolonged times to extubation of 35, 25, 20, 15, and 10%, respectively. The rank distribution of the anesthesiologist corresponding to the adjusted incidence of 25% is skewed to the left. This indicates that this particular anesthesiologist had greater incidence of prolonged times to extubations compared with other anesthesiologists. Two anesthesiologists corresponding to 15 and 20% incidences have wider distributions at the middle of the range. The rank distribution of the anesthesiologist with the lowest incidence of prolonged times to extubation (10%) is skewed to the right. Note that the upper range of the y-axis for *A* and *B* is different, 3 and 25, respectively. The upper limits for the other figures go up to 81.

that these efforts achieve the desired effect. The metric itself may be inappropriate for the proposed purpose. The metric may not actually be under the control of the accountable provider. Providers or groups may be inappropriately identified as failing to meet a metric although critical covariates were not taken into consideration. In addition, basic statistical issues may be ignored, and an individual identified as an outlier when, in fact, the individual's performance is not significantly different from others.

From an operational perspective, focusing on the performance of outlying providers, while seeming to be "intuitively obvious," may not actually result in any meaningful changes in overall performance, thereby potentially wasting effort and resources.

Prolonged time to extubation is one such proposed individual or group metric. However, its use shares the caveats noted above. Simply defining a provider as being an outlier by setting a fixed boundary and failing to incorporate important covariates in the analysis can result in an unreasonably large fraction of providers being defined as outliers. For example, an anesthesiologist who cares principally for patients undergoing long-duration procedures in the prone position might inappropriately be declared as an outlier. Another individual who cares principally for patients undergoing brief procedures in the supine position with a lower incidence of prolonged extubations might be overlooked.

In this study, we showed that prolonged times to tracheal extubation can reliably and validly[12] be monitored by using Bayesian methods incorporating important covariates and key statistical factors (*e.g.*, the impact of widely differing case numbers between providers). Clinically unusual care can be identified and an individual's clinical practice from a quality management perspective reviewed. However, while identifying a provider whose performance is deemed "outlying" may have value from an individual quality assessment perspective, and from a lifelong learning perspective of the faculty anesthesiologist, we also demonstrated that targeting such providers for improvement would not have meaningful impact on OR workflow. The few (*e.g.*, 1 or 2) true outliers

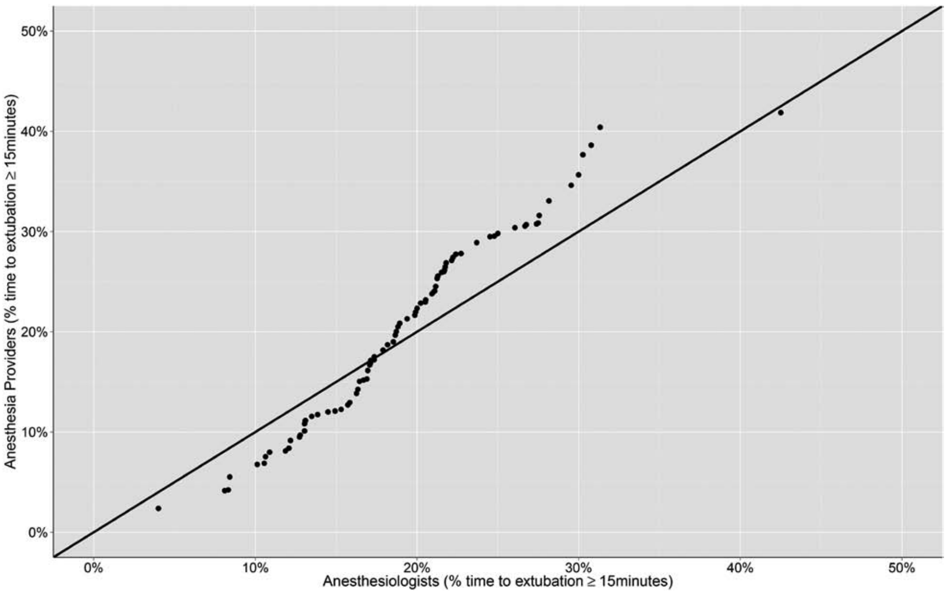

Fig. 8. The quantile-quantile plot comparing anesthesiologists *versus* the anesthesia providers adjusted to a representative patient. This is a plot of sorted quantiles for anesthesiologists against the sorted quantiles for anesthesia providers. When the two distributions are similar, the plot should lie on the identity line (*the straight line*). Below the 20% incidence of prolonged extubation, anesthesia providers have lower incidence compared with anesthesiologists. Above 20%, incidences for anesthesia providers exceed the incidences for anesthesiologists. Above 20%, except the anesthesiologist with higher incidence of prolonged times to extubations (anesthesiologist no. 35), all the anesthesiologists have values on the upper side of the identity line indicating a location shift. In addition, we can see from this plot that incidence ranges differ. Except the outlier anesthesiologist, the upper range of incidence for the anesthesiologists is less than 35%. In contrast, there are several anesthesia providers with incidences above 35%. This plot visually displays our conclusion of larger variability among anesthesia providers compared with anesthesiologists.

each account for too few of all of the prolonged times to extubations.

Previously, we showed that prolonged times to extubations significantly and substantively reduce OR workflow (see introductory text of the current paper).[1–7] When they occur, members of the OR team are idle waiting for extubation (*i.e.*, prolonged extubations are bottlenecks to patient flow).[2] The current results show that change needs to be made by reducing the *average* behavior of everyone, not just a selected few. This might be accomplished administratively (*e.g.*, by altering agents available to providers)[1,6] or by daily case tracking and feedback to any providers whose times were prolonged after accounting for key covariates. A process similar to this has been shown to be effective in reducing operative fresh gas flows (and hence volatile agent utilization).[25,26] The increased variability among care providers (assuming they were responsible for the conduct of the entire anesthetic) is not surprising in a university center with trainees of varying degrees of experience. However, because the variability in the incidence of prolonged extubations was greater among anesthesia providers than anesthesiologists ($P < 0.00001$, WMWodds > 1,000), providing feedback to anesthesiologists should be targeted to their roles as managers ("supervisors"), less so as clinicians directly influencing the extubation times during patient care.[27]

The finding of lack of managerial (economic) value in comparing individual anesthesiologists and anesthesia providers based on a clinically collected measure that combine clinical and operational features matched that which has been found for other endpoints. For example, anesthesiologists differ substantively in their patients' initial postanesthesia care unit pain scores on univariate analysis.[28] These differences disappear when controlled for the two principal predictors, patient age and the nurse obtaining the pain score. Similarly, anesthesiologists differ substantively in their odds of a patient complaint by univariate analysis,[29] but again this disappears when controlled for patient age and tardiness of case start time. In other words, apparent differences among providers can be unrelated to the specific actions of those providers and "targeting" those providers would be fruitless. In contrast, anesthesiologists differ significantly in their clinical care, teaching, and teamwork when evaluated by the anesthesia providers with whom they work.[30–34]

The principal limitation of our study results is that they are from just one university hospital, and several features of that hospital are prominent. First, nearly all cases were performed with an anesthesiologist supervising an anesthesia provider (*e.g.*, resident physician or nurse anesthetist). Our findings of lack of value in monitoring for outlier individuals would likely not apply to a department at which anesthesiologists personally deliver most anesthetics.

In contrast, because we have studied a hospital with nurse anesthetists, the results for anesthesia providers would only apply where there are anesthetic team model. We have no insight as to whether the variability among anesthesiologists would more closely mimic that of our anesthesiologists at a hospital with only anesthesiologists or would more closely reflect our anesthesia providers. Second, because there are so many different anesthesiologists and anesthesia providers, we could not analyze interactions[32]; this would be a good focus for future research. Third, individuals present at the time of extubation may or may not have been responsible for the majority of the conduct of the anesthetic. Fourth, our department's overall incidence of prolonged times to tracheal extubation was 20%, and even 18% among the patients with the lowest risk. In contrast, the incidences were 15 and 15% at two previously studied hospitals,[1,2,35] and 7% in the original (phase IV study) use of propofol.[3] Our department has a low use of the fastest drug,[1,6,36] desflurane, but given the homogeneity of the large incidence among all types of patients, this cannot explain the observation.

Quantitative neuromuscular monitoring is used for all general anesthetics,[37] unlike at the other facilities,[1–3] and with sugammadex not available, perhaps extubation times are often limited by waiting for reversal. Regardless, such a systematic effect would not influence our conclusions.

In conclusion, prolonged extubation is an important metric, but that is minimally influenced by differences among individual practitioners. Monitoring few outlier anesthesiologists (or anesthesia providers) with unusually great incidences of prolonged times to extubations is not warranted. If change is desired, monitoring and feedback to groups of individuals needs to focus on anesthesia providers.

Acknowledgments

The authors thank Gregory Hopson, B.A., M.I.S., from the Department of Anesthesia, University of Iowa, Iowa City, Iowa, for his assistance with extracting the data from the anesthesia information management systems for this study.

Competing Interests

The authors declare no competing interests.

Correspondence

Address correspondence to Dr. Bayman: University of Iowa Hospitals and Clinics, 6439 JCP, 200 Hawkins Drive, Iowa City, Iowa. emine-bayman@uiowa.edu. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. Anesthesiology's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

Appendix

Details of the Bayesian Model and the Weighted Average Calculation

Methods are described in terms of anesthesiologists in this section. Same methods can be applied to anesthesia providers everywhere anesthesiologists were referred.

A.1. The Bayesian Model

Let n_k denote the number of anesthetics with extubation by anesthesiologist k (k = 1, …, K), where K is the current number of anesthesiologists (or anesthesia providers) in the department from January 1, 2012 to December 31, 2013. $y_{ik} = 1$ denotes not having a prolonged time to extubation. In other words, $y_{ik} = 1$ if the time from dressing on the patient until extubation of the trachea was less than 15 min for anesthetic i ($i = 1,\ldots, n_k$) for anesthesiologist k; and $y_{ik} = 0$ if the duration was 15 min or longer. Assuming each anesthesiologist's incidence of prolonged endotracheal extubations is independent of another anesthesiologist, y_{ik} are Bernoulli random variables, and the probability of not having a prolonged time to extubation can be denoted by p_{ik}. In other words,

$$y_{ik} \mid p_{ik} \sim \mathrm{Bin}\left(n_k, p_{ik}\right).$$

The logit link is used to normalize the incidence of prolonged time to extubation. The log odds of a prolonged time to extubation for anesthetic i with anesthesiologist k is denoted as follows:

$$\theta_{ik} = \mathrm{logit}\left(p_{ik}\right) = \ln\left[p_{ik} / \left(1 - p_{ik}\right)\right].$$

θ_{ik} can be written as a function of characteristics of the patient and the surgical procedure. For example, the final model with the significant covariates can be written as follows:

$$\begin{aligned}\theta_{ik} &= \mu + \beta_{\mathrm{Pr}}\mathrm{Prone} + \beta_{\mathrm{BU}}\mathrm{BaseUnit} + \beta_{\mathrm{Dr}}\mathrm{DressTime}\\ &\quad + \beta_{\mathrm{PrBU}}\mathrm{Prone} \times \mathrm{BaseUnit} + \delta_k,\end{aligned}$$

where μ is the intercept in the logit scale, β_{Pr}, β_{BU}, β_{Dr}, and β_{PrBU} are coefficients for the independent covariates and δ_k represents the random anesthesiologist effect. It should be noted that these parameters were defined on the logit scale. Prone is 1 when the patient's positioning is prone and 0 otherwise; base unit is 1 when the numbers of American Society of Anesthesiologists' Relative Value Guide base units are greater than or equal to 11 units and 0 otherwise. The time from operating room (OR) entrance until the dressing has been placed was divided to two branches in the decision tree (fig. 1). However, this variable was used as a continuous variable in the actual analyses. Two-times square-root transformation provided the closest to linear relation with the logit probability of prolonged tracheal extubation for this continuous variable. Therefore, DressTime is the 2 × sqrt (time from OR entrance until the dressing has been placed).

Under the exchangeability assumption, anesthesiologists are considered to be sampled from a common distribution, namely a Normal distribution with mean of 0 and SD of σ. In mathematical notation, this can be written as follows: δ_k ~ Normal (0, σ^2).

A prior distribution is "weakly informative" if it is set up so that the information it provides is intentionally weaker than the available prior knowledge. Weakly informative prior distributions were used for the overall mean, μ, and the coefficients for the fixed effects, β_{Pr}, β_{BU}, β_{Dr}, and β_{PrBU}. Namely, the prior distribution for the overall mean was assumed to have a normal distribution with mean 0 and SD 2; $\mu \sim \text{Normal}\left(0, 2^2\right)$. Using units in the probability scale, the 95% CI of the overall mean according to this prior distribution ranges between 2% (inverse logit [0 − 1.96 × 2]) and 98% (inverse logit [0 + 1.96 × 2]).

Similarly, prior distributions for binary covariates (Prone and Base Unit 11 or greater units *vs.* less than 11 units) were assumed to be normal distribution with mean 0 and SD of 2, $\beta_i \sim \text{Normal}(0, 2^2)$.

Because time from OR entrance until the dressing has been placed with the two-times square-root transformation was on the continuous scale and has a wider scale, the SD of this prior normal distribution was assumed to be 0.1, $\beta_{Dr} \sim \text{Normal}\left(0, 0.1^2\right)$. The same prior distribution was used for the interaction term: $\beta_{PrBU} \sim \text{Normal}\left(0, 0.1^2\right)$.

When the prior distribution includes the available prior knowledge, it is called an "informative" prior distribution. An informative inverse-gamma prior distribution was used for the between-anesthesiologist variance of the probability of prolonged time to extubation, $\sigma^2 : \sigma^2 \sim \text{Inv} - \text{Gamma}\ (\alpha, \beta)$, with mean 0.125 and the SD 0.05. For this inverse-gamma distribution, the prior probability is 95% that any anesthesiologist's log odds of prolonged tracheal extubation lies between 33 and 67%. When the intercept term is 0 (μ = 0), the 95% CI for the time to tracheal extubation being shorter than 15 min is 0.2 to 99.8% for this inverse-gamma distribution. For more details on how to calculate the 95% CI, see the dissertation by Bayman.[17] Values of σ close to 0 represent greater homogeneity of anesthesiologists. As a sensitivity analysis, a noninformative inverse-gamma prior distribution was also used for the between-anesthesiologist variance, $\sigma^2 \sim \text{Inv} - \text{Gamma}\,(0.001,\ 0.001)$ for adjusted models of the anesthesiologists and anesthesia providers.

A.2. Individual Probability

Chaloner and Brant defined outlier as an observation with a large random error. The *k*th anesthesiologist is defined as an outlier, in a linear model with normally distributed random errors, ε_i, with mean 0 and variance σ^2, if $|\varepsilon_i| > m\sigma$ for some *m*. The choice of *m* can be chosen to reflect the fact that the prior probability of observing an outlier is small. Given that the distributions of ε_i's are independent and normally distributed with mean 0, the prior probability of the *k*th anesthesiologist being an outlier can be written as follows. The prior probability that the *k*th anesthesiologist has a significantly greater (or lower) incidence of prolonged tracheal extubations than the other anesthesiologists:

$$\Pr\left(|\ \varepsilon_i| > m\right) = \Pr\left(\varepsilon_i > m\right) + \Pr\left(\varepsilon_i < -m\right) = 2 \times \Phi\left(-m\right),$$

where Φ (*z*) is the standard normal distribution function. The prior probability that the *k*th anesthesiologist is NOT an outlier equals: 1 − 2 × Φ (−*m*).

The prior probability of an anesthesiologist being an outlier can be modeled based on the overall (departmental) probability or the individual probability (see Bayesian Outlier Detection Methods). For the individual probability situation, the probability of each anesthesiologist being an outlier was set to 5%. In this circumstance, *m* becomes 1.96.

A.3. Overall Probability

As a sensitivity analysis, the prior probability of each anesthesiologist being an outlier can be calculated from departmental norms. As explained in the above section, the prior probability that the *k*th anesthesiologist is an outlier can be written as:

$$\Pr\left(|\ \varepsilon_i| > m\right) = \Pr\left(\varepsilon_i > m\right) + \Pr\left(\varepsilon_i < -m\right) = 2 \times \Phi\left(-m\right).$$

The prior probability that the *k*th anesthesiologist is NOT an outlier equals: 1 − 2 × Φ (−*m*).

There are a total of *K* anesthesiologists. Under the independence assumption of the random error terms, the prior probability that none of the *K* anesthesiologists is an outlier can be written as follows:

$$\left[1 - 2\Phi(-m)\right]^K .$$

The prior probability of not detecting any anesthesiologist in the department, during the 2 yr, having an outlier incidence of prolonged time to extubation should be high and is set to 95%. In other words,

$$\left[1 - 2\Phi(-m)\right]^K = 0.95.$$

This corresponds to the probability of "at least one anesthesiologist in the department during the 2-yr study period having a significantly greater incidence of prolonged time to extubations than the other anesthesiologists" being equal to 5%.

The prior probability of one specific anesthesiologist *k* being an outlier, when there are *K* anesthesiologists, is 2Φ (−*m*), where,

$$m = \Phi^{-1}\left[0.5 + \tfrac{1}{2} \times \left(0.95^{1/K}\right)\right].$$

In our data set for 2 yr, there were 81 anesthesiologists: *K* = 81. Thus, *m* = 3.42 and the prior probability of each anesthesiologist having an outlier incidence of prolonged time to extubations is 0.0006. Similarly, there were 116 anesthesia providers. Therefore, *K* = 116, *m* = 3.51, and the prior probability of each anesthesia provider having an outlier incidence of prolonged time to extubations is 0.00044.

A.4. Sensitivity Analyses

For the original calculations, it was assumed that the between-anesthesiologist variance in the logit scale has an inverse-gamma distribution with parameters $\alpha = 9$ and $\beta = 1$. As a sensitivity analysis, a noninformative inverse-gamma prior distribution was used; specifically inverse-gamma ($\alpha = 0.001$, $\beta = 0.001$) as recommended by Spiegelhalter *et al.*[14]

A different prior distribution for the overall mean, μ, was also examined; $\mu \sim$ uniform $(-4.595, 4.595)$ instead of $\mu \sim$ normal $(0, 2^2)$. The justification for the uniform prior distribution was as follows. The probability of observing a prolonged time to extubation varies between 0 and 1. If we define this probability between 0.01 and 0.99, the logit of this range correspond to −4.595 to 4.595.

When calculations for anesthesiologists and anesthesia providers for the adjusted models were repeated for these different sets of prior probabilities, same individuals were identified as outliers. This shows that our results are insensitive to these two choices of prior distributions of between-anesthesiologist variance.

A.5. WinBUGS Model for Anesthesiologists, Adjusted Model

```
model
   {
A    for (i in 1:27757){
B    goodoutPE[i] ~ dbern(p[i])
C    logit(p[i]) <- theta[i]
D    theta[i] <- mu + beta_Prone*Prone[i] + beta_BU*
     BUGr11[i] + beta_Dr * ORDr[i] + beta_PrBU*
     Prone[i] * BUGr11[i] +delta[anes[i]]
   }
E       for(k in 1:81){
F       Post.delta.3.42[k] <- step(delta[k] - 3.42*sigma.e)
   + step(-delta[k] - 3.42*sigma.e)
   #m <- qnorm(0.5 + 0.95^(1/81)/2) = 3.42
   }
G    Prob.sum <- sum(Post.delta.3.42[])
H    Prob.any.g3.42 <- step(Prob.sum -1)
   for(j in 1:81){
I    delta[j] ~ dnorm(0, prec.delta)
   }
J    mu~ dnorm(0, 0.25) # sd = 2
K    beta_Prone~ dnorm(0, 0.25) # sd = 2
L    beta_BU~ dnorm(0, 0.25) # sd = 2
M    beta_Dr~ dnorm(0, 100) # sd = 0.1
N    beta_PrBU~ dnorm(0, 100) # sd = 0.1
O    prec.delta ~ dgamma(9, 1)
P    sd.delta <- 1/sqrt(prec.delta)
   }
```

At **A**, *i* refers to the *i*th of 27,757 anesthetics performed in 2 yr.

At **B**, goodout PE[*i*] is a binary variable denoted by 1 if the time to tracheal extubation is shorter than 15 min and 0 otherwise. The incidence of time to extubation being shorter than 15 min has a Bernoulli distribution with the probability of p[*i*].

At **C**, the logit transformation is applied to the incidence of time to extubation being shorter than 15 min, similar to the logistic regression model.

At **D**, the logit probability of the incidence of time to tracheal extubation being less than 15 min is written as a function of overall intercept (μ), patients' position (prone *vs.* not prone), the numbers of American Society of Anesthesiologists' Relative Value Guide base unit (11 or greater *vs.* less than 11), time from OR entrance until the dressing has been placed (after two times the square-root transformation), and the random anesthesiologist effect (delta[anes(*i*)]).

At **E**, *k* stands for the *k*th of 81 anesthesiologists that was assessed for performance within 2 yr.

At **F**, the posterior probability of being outlier for anesthesiologist *k* (Post.delta.3.42[*k*]) was calculated. This probability was calculated based on the number of anesthesiologists compared and is 81 for this example. First, the prior probability of being outlier, *m*, was calculated (see appendix A.3). The step function was used to calculate how many times the random anesthesiologist effect was more extreme than *m*. Step(e) returns 1 if $e \geq 0$ and 0 otherwise. When the prior probability for each individual set to 5% (see appendix A.3), instead of 3.42, 1.96 was used.

At **G**, the sum of posterior probabilities of being outlier for all anesthesiologists was calculated.

At **H**, the posterior probability of at least one anesthesiologist being an outlier (Prob.any.g3.42) was calculated.

At **I**, a random normal distribution was defined as a prior distribution for the random provider effect. WinBUGS used the mean and precision to indicate the normal distribution. The normal distribution was centered at mean of 0 and the SD is $1/\sqrt{\text{prec.delta}}$.

At **J**, a prior distribution for the intercept term (μ) was defined as a normal distribution with mean 0 and SD of 2.

At **K**, a prior distribution was defined for the slope of patient's prone position. This is a normal distribution with mean 0 and SD of 2 and therefore a weak informative prior distribution.

At **L**, a prior distribution was defined for the slope of the numbers of American Society of Anesthesiologists' Relative Value Guide base unit of the anesthetic (11 or greater *vs.*

less than 11). This is a normal distribution with mean 0 and SD of 2.

At **M**, a prior distribution was defined for the slope of two times the square-root transformed (Box–Cox transformed) time from OR entrance until the dressing has been placed. This is a normal distribution with mean 0 and SD of 0.1.

At **N**, a prior distribution was defined for the slope of the interaction of Prone and the two times the square-root transformed (Box–Cox transformed) time from OR entrance until the dressing has been placed. This is a normal distribution with mean 0 and SD of 0.1.

At **O**, as usual for the precision of the normal distribution (see WinBUGS manual), the prior distribution for the precision of the random provider effect was defined as a gamma distribution with parameters 9 and 1. The mean of this gamma distribution is 9 and it's SD is 3.

At **P**, the conversion between the SD (sd.delta) and the precision (prec.delta) was defined.

A.6. Details of the Weighted Average Calculation

The variance of the posterior mean was obtained from the Bayesian analyses results. The inverse of the variance for each anesthesiologist was used as the weight for that anesthesiologist's random effect. Therefore, a greater weight was given to the random effects for those anesthesiologists with greater number of extubations.

References

1. Dexter F, Bayman EO, Epstein RH: Statistical modeling of average and variability of time to extubation for meta-analysis comparing desflurane to sevoflurane. Anesth Analg 2010; 110:570–80
2. Masursky D, Dexter F, Kwakye MO, Smallman B: Measure to quantify the influence of time from end of surgery to tracheal extubation on operating room workflow. Anesth Analg 2012; 115:402–6
3. Apfelbaum JL, Grasela TH, Hug CC Jr, McLeskey CH, Nahrwold ML, Roizen MF, Stanley TH, Thisted RA, Walawander CA, White PF: The initial clinical experience of 1819 physicians in maintaining anesthesia with propofol: Characteristics associated with prolonged time to awakening. Anesth Analg 1993; 77(4 suppl):S10–4
4. Dexter F, Epstein RH: Increased mean time from end of surgery to operating room exit in a historical cohort of cases with prolonged time to extubation. Anesth Analg 2013; 117:1453–9
5. Epstein RH, Dexter F, Brull SJ: Cohort study of cases with prolonged tracheal extubation times to examine the relationship with duration of workday. Can J Anaesth 2013; 60:1070–6
6. Agoliati A, Dexter F, Lok J, Masursky D, Sarwar MF, Stuart SB, Bayman EO, Epstein RH: Meta-analysis of average and variability of time to extubation comparing isoflurane with desflurane or isoflurane with sevoflurane. Anesth Analg 2010; 110:1433–9
7. Vitez TS, Macario A: Setting performance standards for an anesthesia department. J Clin Anesth 1998; 10:166–75
8. Bayman EO, Dexter F, Todd MM: Assessing and comparing anesthesiologists' performance on mandated metrics using a Bayesian approach. Anesthesiology 2015; 123:101–15
9. de Ville B, Neville P: Decision Trees for Analytics Using SAS Enterprise Miner. Cary, North Carolina, SAS Institute, 2013
10. Ehrenfeld JM, Henneman JP, Peterfreund RA, Sheehan TD, Xue F, Spring S, Sandberg WS: Ongoing professional performance evaluation (OPPE) using automatically captured electronic anesthesia data. Jt Comm J Qual Patient Saf 2012; 38:73–80
11. Chaloner K, Brant R: A Bayesian approach to outlier detection and residual analysis. Biometrika 1988; 75:651–9
12. Bayman EO, Chaloner KM, Hindman BJ, Todd MM; IHAST Investigators: Bayesian methods to determine performance differences and to quantify variability among centers in multi-center trials: The IHAST trial. BMC Med Res Methodol 2013; 13:5
13. Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Anal 2006; 1:1–19
14. Spiegelhalter DJ, Abrams KR, Myles JP: Bayesian Approaches to Clinical Trials and Health-care Evaluation. Chichester, United Kingdom, John Wiley & Sons, 2004
15. Bayman EO, Chaloner K, Cowles MK: Detecting qualitative interaction: A Bayesian approach. Stat Med 2010; 29:455–63
16. Kass RE, Raftery AE: Bayes factors. J Am Stat Assoc 1995; 90:773–95
17. Bayman EO: Bayesian Hierarchical Models for Multicenter Clinical Trials: Power and Subgroup Analyses. Iowa, Department of Biostatistics, University of Iowa, 2008, p 205
18. Lunn DJ, Thomas A, Best N, Spiegelhalter D: WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. Stat Comput 2000; 10:325–37
19. R Development Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, 2013
20. Brooks SP, Gelman A: General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998; 7:434–55
21. Divine G, Norton HJ, Hunt R, Dienemann J: Statistical grand rounds: A review of analysis and sample size calculation considerations for Wilcoxon tests. Anesth Analg 2013; 117:699–710
22. Dexter F: Wilcoxon-Mann-Whitney test used for data that are not normally distributed. Anesth Analg 2013; 117:537–8
23. Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA: Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. J Clin Epidemiol 2005; 58:261–8
24. Gelman A, Hill J, Yajima M: Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 2012; 5:189–211
25. Dexter F, Maguire D, Epstein RH: Observational study of anaesthetists' fresh gas flow rates during anaesthesia with desflurane, isoflurane and sevoflurane. Anaesth Intensive Care 2011; 39:460–4
26. Epstein RH, Dexter F, Patel N: Influencing anesthesia provider behavior using anesthesia information management system data for near real-time alerts and *post hoc* reports. Anesth Analg 2015; 121:678–92
27. Dexter F, Wachtel RE: Strategies for net cost reductions with the expanded role and expertise of anesthesiologists in the perioperative surgical home. Anesth Analg 2014; 118:1062–71
28. Wanderer JP, Shi Y, Schildcrout JS, Ehrenfeld JM, Epstein RH: Supervising anesthesiologists cannot be effectively compared according to their patients' postanesthesia care unit admission pain scores. Anesth Analg 2015; 120:923–32
29. Kynes JM, Schildcrout JS, Hickson GB, Pichert JW, Han X, Ehrenfeld JM, Westlake MW, Catron T, Jacques PS: An analysis

of risk factors for patient complaints about ambulatory anesthesiology care. Anesth Analg 2013; 116:1325–32

30. de Oliveira Filho GR, Dal Mago AJ, Garcia JH, Goldschmidt R: An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. Anesth Analg 2008; 107:1316–22
31. Hindman BJ, Dexter F, Kreiter CD, Wachtel RE: Determinants, associations, and psychometric properties of resident assessments of anesthesiologist operating room supervision. Anesth Analg 2013; 116:1342–51
32. Dexter F, Ledolter J, Smith TC, Griffiths D, Hindman BJ: Influence of provider type (nurse anesthetist or resident physician), staff assignments, and other covariates on daily evaluations of anesthesiologists' quality of supervision. Anesth Analg 2014; 119:670–8
33. Dexter F, Ledolter J, Hindman BJ: Bernoulli Cumulative Sum (CUSUM) control charts for monitoring of anesthesiologists' performance in supervising anesthesia residents and nurse anesthetists. Anesth Analg 2014; 119:679–85
34. Dexter F, Masursky D, Hindman BJ: Reliability and validity of the anesthesiologist supervision instrument when certified registered nurse anesthetists provide scores. Anesth Analg 2015; 120:214–9
35. Wachtel RE, Dexter F, Epstein RH, Ledolter J: Meta-analysis of desflurane and propofol average times and variability in times to extubation and following commands. Can J Anaesth 2011; 58:714–24
36. Todd MM, Hindman BJ, King BJ: The implementation of quantitative electromyographic neuromuscular monitoring in an academic anesthesia department. Anesth Analg 2014; 119:323–31