Assessing and Comparing Anesthesiologists' Performance on Mandated Metrics Using a Bayesian Approach

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

ABSTRACT

Background: Periodic assessment of performance by anesthesiologists is required by The Joint Commission Ongoing Professional Performance Evaluation program.

Methods: The metrics used in this study were the (1) measurement of blood pressure and (2) oxygen saturation (Spo_2) either before or less than 5 min after anesthesia induction. Noncompliance was defined as no measurement within this time interval. The authors assessed the frequency of noncompliance using information from 63,913 cases drawn from the anesthesia information management system. To adjust for differences in patient and procedural characteristics, 135 preoperative variables were analyzed with decision trees. The retained covariate for the blood pressure metric was patient's age and, for Spo_2 metric, was American Society of Anesthesiologist's physical status, whether the patient was coming from an intensive care unit, and whether induction occurred within 5 min of the start of the scheduled workday. A Bayesian hierarchical model, designed to identify anesthesiologists as "performance outliers," *after* adjustment for covariates, was developed and was compared with frequentist methods.

Results: The global incidences of noncompliance (with frequentist 95% CI) were 5.35% (5.17 to 5.53%) for blood pressure and 1.22% (1.14 to 1.30%) for Spo_2 metrics. By using unadjusted rates and frequentist statistics, it was found that up to 43% of anesthesiologists would be deemed noncompliant for the blood pressure metric and 70% of anesthesiologists for the Spo_2 metric. By using Bayesian analyses with covariate adjustment, only 2.44% (1.28 to 3.60%) and 0.00% of the anesthesiologists would be deemed "noncompliant" for blood pressure and Spo_2, respectively.

Conclusion: Bayesian hierarchical multivariate methodology with covariate adjustment is better suited to faculty monitoring than the nonhierarchical frequentist approach. **(Anesthesiology 2015; 123:101-15)**

THE Joint Commission, a United States–based hospital accreditation organization, requires that all licensed practitioners (*e.g.*, anesthesiologists) undergo periodic Ongoing Professional Practice Evaluation (OPPE). Evaluations must be based, at least in part, on measures of clinical performance and identify providers by name. Results are reported to the hospital.

There are two separate issues when comparing the performance of anesthesiologists. One is the determination of the mean risk-adjusted incidence of their noncompliance with a chosen metric. This would be appropriate for making a comparison between groups and hospitals.[1,2] The second is the determination of outlying individuals *within* the same department. The two issues happen at different hierarchical levels. The OPPE requirement addresses the second issue.

What We Already Know about This Topic

- Although periodic assessment of anesthesiologists is required by some regulatory agencies in the world, there are no broadly accepted quality or safety performance metrics in anesthesia

What This Article Tells Us That Is New

- Noncompliance with simple blood pressure and oxyhemoglobin saturation metrics in approximately 70,000 cases at the University of Iowa was present in up to 43 and 70% of anesthesiologists, respectively, by using frequentist statistics compared with 2.4 and 0%, respectively, by using a Bayesian approach

Anesthesiologists work in different subspecialty areas and perform widely differing numbers of cases. A metric applicable to one subset of anesthesiologists within a department might be meaningless when applied to another. For example,

This article is featured in "This Month in Anesthesiology," page 1A. 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). This study was presented at the 2013 American Society of Anesthesiologists meeting in San Francisco, California, on October 13, 2013, as one of the best clinical abstracts.

Submitted for publication September 11, 2014. Accepted for publication February 22, 2015. From the Departments of Anesthesia and Biostatistics (E.O.B.), Division of Management Consulting, Department of Anesthesia (F.D.), and Department of Anesthesia (M.M.T.), University of Iowa, Iowa City, Iowa.

 Anesthesiology 2015; 123:101-15

the incidence of perioperative mortality might be a meaningful outcome measure for anesthesiologists doing cardiac surgery, but not for anesthesiologists' performing sedation for gastroenterology. Although Haller *et al.*[1] reviewed 108 potential measures, they did not assess their utility for the OPPE process.

Not only is finding a "one-size fits all" metric difficult, so is finding valid analytic methods that avoid the "spurious outlier" problem inherent in the face of anesthesiologist-by-anesthesiologist, patient and procedural, variation. Using raw measures of noncompliance (*e.g.*, percentage of a given clinician's anesthetics not meeting the chosen metric) in conjunction with typical frequentist statistics (*e.g.*, chi-square test or funnel plots) may yield misleading comparisons.

Ehrenfeld *et al.*[3] published two OPPE metrics: measurement of blood pressure before induction and use of end-tidal carbon dioxide monitoring. They monitored anesthesiologists and calculated upper one-sided 95% CIs of the overall incidences of noncompliance as thresholds. Any anesthesiologist with an incidence of noncompliance above the threshold was classified as "not passing the metric." However, they did not make adjustments for patient and procedural covariates. They also excluded pediatric cases and anesthesiologists who supervised less than 60 anesthetics during the assessment period, inconsistent with The Joint Commission rule to include all providers.

We developed a Bayesian hierarchical model to identify "performance outliers" after adjustment for covariates. The inclusion of relevant covariates eliminates the need to exclude large numbers of cases (*e.g.*, pediatrics). The Bayesian approach is more statistically powerful than the frequentist methods when the sample sizes for individual participants are heterogeneous and when some are small.[4] Therefore, no providers are excluded based on their numbers of anesthetics.

The goal of this study was to compare the characteristics and results of different statistical methods used to detect those anesthesiologists who might be judged as "outliers," with and without efforts to take into account different patient and subspecialty characteristics. We evaluated the influence of risk adjustment for each metric on results and compared our Bayesian methods to Ehrenfeld's frequentist-observed percentage with no covariate adjustment.

Materials and Methods

The Human Subject Research Determination form submitted to the University of Iowa Institutional Review Board (Iowa City, Iowa) was determined that this retrospective quality assurance project concerns primarily clinical activities and does not meet the regulatory definition of human subjects research (see table 1, Supplemental Digital Content 1, http://links.lww.com/ALN/B145, for the details of the structured query language logic to create the analyzed dataset).

We began with information from 79,327 anesthetics extracted from the University of Iowa's Epic (Epic Systems, Inc., USA) anesthesia information management system (AIMS) (appendix, Anesthesia Medical System Time Stamps), starting shortly after Epic "Go-Live" (November 16, 2010) and extending through June 30, 2013. We then focused on general anesthetics that were initiated with one of the five different agents: propofol, etomidate, sevoflurane, desflurane, and rocuronium. Isoflurane was not included because for no case was it the first agent used (*i.e.*, its use was always preceded by one of the above noted medications). Desflurane and rocuronium were included because the study included patients who were already intubated and sedated and transported from an intensive care unit (ICU) to the operating rooms (ORs), and these two agents were sometimes the first given in the OR. This resulted in a dataset of 68,220 cases (see table 1, Supplemental Digital Content 1, http://links.lww.com/ALN/B145, which lists the structured query language logic to obtain 68,220 cases). Preoperative and procedural information from these cases were used to develop covariates for subsequent analyses (see sections Blood Pressure Metric—Selection of Covariates and pulse oximetry measured oxygen saturation (Spo_2) Metric—Selection of Covariates). For performance assessments, five sequential, 6-month periods were used, beginning with January 2011 through June 2011 and ending with January 2013 through June 2013. The 6-month periods were used because this is an accepted interval for periodic OPPE assessments (assessments must be more frequent than once-per-year).

Our Bayesian method works fully for all providers (*i.e.*, for our local quality improvement, all providers are included). However, in our analyses for this article, in order for an anesthesiologist to be included in the performance assessment for a period, the anesthesiologist being assessed had to be working for the department during that entire 6-month period. We did this to provide scientific results in this article that are generalizable to other institutions. The number of anesthesiologists doing hardly any cases for a 6-month period would not apply elsewhere. The effect of this was to reduce the sample size for performance analyses to 63,913 general anesthetics.

To assess and compare the performances of anesthesiologists, two metrics were chosen. The criteria in choosing the metrics were that each should (1) apply to all anesthesiologists performing general anesthetics in the department; (2) be measured objectively; and (3) be present in our Epic AIMS to permit ready extraction every 6 months. The metrics chosen were (1) the time of the first recorded arterial or noninvasive blood pressure (NIBP) in relation to the time of induction[3,5] and (2) the time of the first recorded Spo_2 in relation to induction (see section Definition of "Time of Induction" for more information). For an anesthesiologist to be deemed compliant, these values needed to be recorded either before or coincident with the first appearance of any of the five aforementioned medications (defined as the start of induction). Because, for more than half of the records, the first induction dose was propofol, and because the time of propofol administration was not recorded automatically

(*i.e.*, it was manually entered and may have been measured with error), we chose to more specifically define the "compliant" interval as extending until less than 5 min *after* first drug administration. If the first blood pressure/Spo_2 recording was 5 min or greater after induction, the anesthesiologist in that case was deemed noncompliant.

Blood Pressure Metric: NIBP or Arterial Blood Pressure

The earliest recording of any NIBP after the start of continuous anesthesia presence (anesthesia start time) was used. If the patient had an arterial blood pressure measurement that preceded the NIBP, then the first arterial blood pressure time after the anesthesia start time was used. Arterial blood pressures were excluded when not physiologically plausible (*e.g.*, flushing the arterial line catheter), defined as systolic blood pressure less than 50 mmHg, systolic greater than 230 mmHg, diastolic less than 30 mmHg, or diastolic greater than 140 mmHg.

Most of the monitors used to measure blood pressure during an anesthetic were recorded automatically in Epic. However, when an ICU patient was transported to the OR, blood pressure may have been displayed on a transport monitor but not recorded in Epic. Ehrenfeld *et al.*[3] addressed this issue by excluding any case for which the patient came from an ICU (personal written communication with Jesse Ehrenfeld, M.D., M.P.H., Associate Professor, Department of Anesthesiology, Vanderbilt University, Nashville, Tennessee, July 25, 2013).[3] We did not do this, for two reasons. First, the incidence of blood pressure noncompliance for patients coming from the ICU (n = 2,274) was not substantially greater than for all other patients (9.45 *vs.* 5.32%), and we wished to include as many patients (and hence providers) as possible in our analysis. Second, and most importantly, this is a systems-based issue with the electronic medical record. Charting is the responsibility of the patient care team, regardless of whether it is recorded automatically or manually. Therefore, rather than deleting these cases, we included the origination of the patient (from ICU) in the adjusted analysis.

Ehrenfeld *et al.*[3] also excluded 35% (46 of 128) of anesthesiologists because they performed fewer than 60 general anesthetics during the studied period. Fewer of our anesthesiologists would have been excluded (4 of 56, 7.1%), but because the Bayesian method is not influenced by the number of anesthetics performed, we included all anesthesiologists regardless of the number of general anesthetics performed during the studied 6-month period, as long as the anesthesiologist worked for the department during that entire 6-month period. This meets the OPPE mandate.

Definition of "Time of Induction"

For intravenous drugs, the time of induction was considered the earliest recorded time of administration after the anesthesia start time.

For propofol, etomidate, and/or rocuronium, an induction dose exceeding chosen thresholds was used. The threshold for propofol was 0.125 mg/kg, chosen to be greater than the typical ICU sedation dose of 25 $\mu g\ kg^{-1}\ min^{-1} \times 5$ min. As this was approximately 20% of a typical induction dose, the (approximate) corresponding criteria applied were 0.06 mg/kg for etomidate and 0.12 mg/kg for rocuronium. The total dose over 5 min was used when there was more than one dose or an infusion. An example is provided in the appendix (Examples of Calculating the Total Dose of Propofol).

For the volatile anesthetics (sevoflurane or desflurane), the first time after the anesthesia start time when the end-tidal percentage concentration automatically exceeded a threshold was recorded. Thresholds used were 0.2 times the minimum alveolar concentration, that is, 0.42% for sevoflurane and 1% for desflurane. The reason for using threshold values for end-tidal concentrations was that residual subhypnotic concentrations from previous anesthetics in the OR on the same day frequently "bled over" into a subsequent case, potentially resulting in an erroneous induction time (appendix, Using Thresholds for End-tidal Concentrations).

Blood Pressure Metric—Selection of Covariates

We started with a dataset consisting of *all* the preoperative characteristics available in the AIMS for the 68,220 patients receiving general anesthesia during the November 2010 through June 2013 period (tables 1–3, Supplemental Digital Content 1, http://links.lww.com/ALN/B145, 135 variables). Classification tree analyses were then performed by using SAS Enterprise Miner 7.1 (SAS Institute, Inc., USA). Models were compared based on the mean squared error. The classification/decision tree analyses created a hierarchy of branches.[6] Each variable was divided into as many as three branches. Each variable was used only once in the decision tree. For these analyses, a single dataset including the data from the entire period was used instead of five datasets from each of the 6-month periods.

The use of age in the model reduced the mean squared error more than the use of any of the other 134 variables (see tables 2 and 3, Supplemental Digital Content 1, http://links.lww.com/ALN/B145). Once age was included, adding none of the other 134 variables meaningfully reduced the mean squared error (reduction on the mean squared error for each other variable ≤0.25%).

The three age categories selected automatically by SAS Enterprise Miner were (1) age less than 7 yr and 3 months; (2) age between 7 yr and 3 months and 12 yr and 9 months; and (3) age 12 yr and 9 months or older (fig. 1). As shown in this figure, blood pressure was not checked within 5 min after induction for 5.32% of the 68,200 general anesthetics. The incidence of noncompliance was the greatest for the youngest age group (*i.e.*, principally for pediatric anesthesiologists) (fig. 2). Because pediatric anesthesiologists often induce anesthesia in their youngest patients with volatile agents before placing monitors, this observation was expected, suggesting that our methods are valid.

The Bayesian method uses logistic regression models. Knowing the model variable from the SAS Enterprise

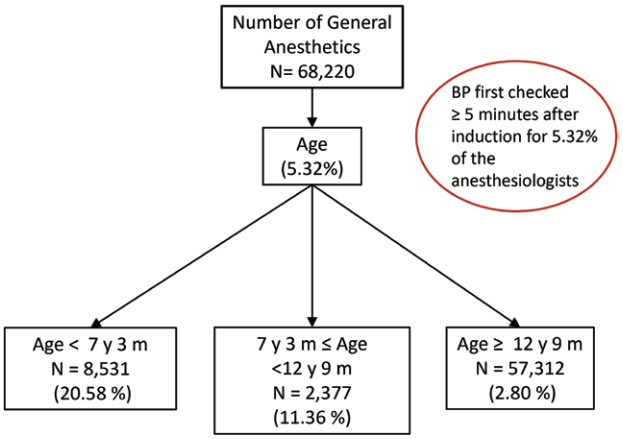

Fig. 1. Decision tree model for the blood pressure (BP) outcome. Ehrenfeld *et al.*[3] retrospectively studied the electronic anesthesia records of patients for the purpose of prolonged blood pressure gaps in anesthesia records. m = months; y = years.

Miner, Box-Cox transformation was used to determine the best transformation for age as a continuous variable to satisfy the assumption of a linear relation between transformed age and the logit. The transformation used was two times the square root of age. The mean squared error for the blood pressure metric with age as the covariate was 0.0468. To evaluate, further, whether any of the other 134 variables individually should also be used for covariate adjustment, logistic regression models were fit, and the increases in the area under the curve *versus* from age alone were calculated. It was verified that the model with age was as good as any other model with any additional variable. The increases in the area under the curve, after the inclusion of other variables in addition to age, were all less than 0.6% absolute value. Since both mean squared error from the classification tree analyses and area under the curve statistics from the logistic regression models indicated that no candidate variable made a meaningful change in the model, patient's age (continuous variable with the square-root transformation) was the only characteristic that was used in the Bayesian model for absence of checking the blood pressure within 5 min of the induction of general anesthesia.

Finally, Bayesian hierarchical generalized linear models were fit, adjusting for the patient's age within each of five 6-month periods. In the Bayesian model, patient's age, along with the random anesthesiologist effect, was included.

Spo_2 Metric—Selection of Covariates

Similar steps were followed for Spo_2. Significant covariates detected by the classification tree analyses were the American Society of Anesthesiologists (ASA) physical status score (1 to 3 *vs.* 4 to 6), whether the patient was coming from the ICU, and whether the case was a first start of the workday, all binary variables (fig. 3). The case was considered a first start when the time from the start of the surgical day to induction was 5 min or less. Note that "the time from the start of the surgical day" was entered as a continuous variable in the analyses for the decision tree. SAS Enterprise Miner broke down this variable as a binary covariate.

There were 35 patients with ASA physical status 6 (brain dead), and they were included in the analyses.

The "From ICU" variable reports if the preceding location before the patient was in an OR was an ICU. Locations considered ICUs were the hospital's cardiovascular ICU, emergency medicine department, medical ICU, neonatal ICU, pediatric ICU, and the surgical and neuroscience ICU.

As displayed in the decision tree of figure 3, patients with ASA physical scores 4, 5, or 6 appear not to have had their Spo_2 checked before or within 5 min after induction more often than for patients with lesser ASA scores (5.40 *vs.* 1.09%, respectively). Similarly, the Spo_2 appears not to have been checked before induction more often for those patients coming from an ICU (7.77 *vs.* 4.43%). The sicker patients (ASA ≥4) coming from the ICU did so on transport monitors and sometimes the anesthesia provider did not enter the information into the electronic medical record (again, as described in section Blood Pressure Metric—Selection of Covariates). These patients were included in our study because these are precisely the patients for whom hypotension and/or hypoxemia may influence patient outcome. However, by including these variables in the risk adjustment, each anesthesiologist's behavior was compared with other anesthesiologists addressing the same systems-based issues.

Frequentist Outlier Detection Methods

The method described in the study by Ehrenfeld *et al.*[3] was used to identify outlier anesthesiologists. Ehrenfeld *et al.*[3] calculated a compliance threshold as an upper 95% CI of the overall incidence of noncompliance—and considered providers whose performance was beyond this CI as being "noncompliant." We calculated this frequentist threshold based on the data from a single dataset including the data from all 2.5 yr and then applied it to each of the five 6-month intervals.

Bayesian Outlier Detection Methods

The method described in the study by Chaloner and Brant[7] and Bayman *et al.*[8] was used to identify outliers. The method[8] was developed to detect outliers among centers in multicenter clinical trials. In the current study, they are applied to detect anesthesiologists with outlier behavior. Each anesthetic was attributed to the single anesthesiologist assigned in the electronic medical record at the time of the induction drug administration. The hierarchical model is especially appropriate here because the anesthetics are nested within the anesthesiologists, and the model takes into account patient and procedure characteristics

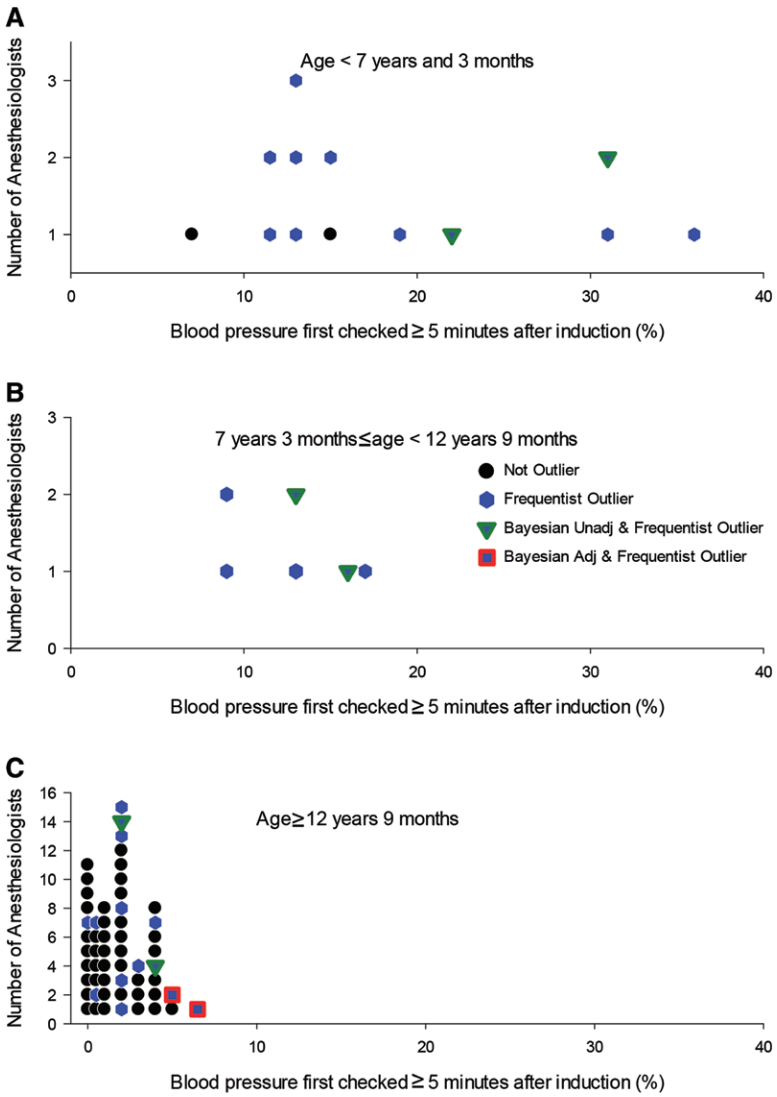

Fig. 2. Dotplot for January 2013 through June 2013 for the blood pressure metric for three age groups. Data from January 2013 through June 2013 for the blood pressure metric are presented in figure 2 for each age group from SAS Enterprise Miner (SAS Institute, Inc., USA; fig. 1). Only those anesthesiologists with 25 or more anesthetics in the 6-month period for each age group were plotted. Therefore, even though there were 57 anesthesiologists studied in this period, the number of *dots* in *A* and *B* would be less than 57. The incidences of noncompliance were 19.96% for the youngest age group (*A*) and 2.13% for those patients in the oldest age group (*C*) in this period. The figure shows that the incidences of not checking (or recording) blood pressure before induction were substantial for the two Bayesian outlier anesthesiologists without risk adjustment (*green triangle*), especially for the middle and oldest age groups. The figure shows also that neither of the two anesthesiologists who was a Bayesian outlier with covariate adjustment (*red squares*) shows up on *A* nor on *B*, indicating that these anesthesiologists performed less than 25 anesthetics for these age groups during the 6-month period. Adj = adjusted; Unadj = unadjusted.

(see sections Blood Pressure Metric—Selection of Covariates and Spo_2 Metric—Selection of Covariates).

Bayesian hierarchical generalized linear models were fit for each metric. Anesthesiologists were assumed to be performing similarly but not identical to one another: exchangeable.[9] In statistical modeling terms, this means that it was assumed that the anesthesiologists' performances were randomly sampled from the same normal distribution. Details of the model are given in the appendix (The Bayesian Model).

In Bayesian analysis, 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.

Two different prior probabilities were examined for an anesthesiologist having a significantly greater incidence of blood pressure (or Spo_2) noncompliance than the other anesthesiologists during the each of five 6-month periods. (1) The prior probability of each anesthesiologist having a significantly greater incidence of blood pressure (or Spo_2) noncompliance than the other anesthesiologists was set to 0.05 (appendix, Individual Prior Probability). (2) The prior probability of *at*

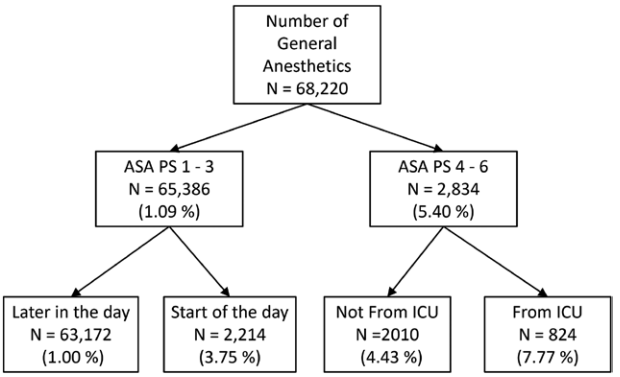

Fig. 3. Decision tree for the pulse oximetry measured oxygen saturation (Spo_2) outcome. Percentages (%) in parentheses indicate the incidences of Spo_2 first checked more than 5 min after the induction. For example, for 65,386 patients with American Society of Anesthesiologist's physical status (ASA PS) 1 to 3, Spo_2 was first checked more than 5 min after the induction 1.09% of the time. Among the anesthetics represented in the figure, there are data for 2,629, 169, and 35 anesthetics with American Society of Anesthesiologists physical status 4, 5, and 6, respectively. Start of the day is a binary variable indicating if the time from the start of the surgical day to induction were ≤5 min *versus* >5 min. At the University of Iowa, the surgical day starts at 8:00 AM on Monday and Tuesday and at 7:15 AM on Wednesday, Thursday, and Friday. The "From ICU" variable reports if the preceding location before the patient was in an operating room was an intensive care unit (ICU). This includes the cardiovascular ICU, medical ICU, neonatal ICU, pediatric ICU, and surgical and neuroscience intensive care.

least one anesthesiologist in the department during each studied 6-month period having a significantly greater incidence of blood pressure (or Spo_2) noncompliance than the other anesthesiologists was set to 0.05. Under the second setting, for 57 anesthesiologists in the department January 2013 through June 2013, the prior probability of each anesthesiologist being an outlier becomes 0.0009 (appendix, Overall Prior Probability).

Prior distributions used for the overall mean and the coefficients for the fixed effects (*i.e.*, patient age and whether the patient is coming from an ICU) were assumed to follow normal distributions as is typical for these types of analyses. The parameters of these prior distributions were set for them to be weakly informative (have very large SDs)[10] (for statistical details and explanations of the WinBUGS model, see the appendix, Adjusted WinBUGS Model for Blood Pressure Outcome - Individual Probability). Analyses were repeated using different prior distributions as sensitivity analyses. Random prior distributions were defined for each anesthesiologist. For each anesthesiologist, posterior probabilities of having a significantly greater incidence of blood pressure (or Spo_2) noncompliance (compared with the other anesthesiologists) were calculated, and the strength of evidence was quantified by the Bayes factor (BF).[9]

The BF is the ratio of the posterior odds in favor of the null to the prior odds of the null.[4] The most common interpretation of BF is that it classifies evidence against the null hypothesis. The evidence is considered "strong," "very strong," and "decisive" when the BF is less than 10^{-1}, $10^{-1.5}$, and 10^{-2}, respectively, according to Jeffrey scale.[9] Kass and Raftery[11] recommend a more conservative interpretation of BF. According to Kass and Raftery, BFs less than 0.33, 0.05, and 0.0067 are classified as "positive," "strong," and "very strong" evidence against the null hypothesis.[11] With both scales, BF greater than 1 provides the evidence for the null hypothesis.

A BF less than 0.1 indicates "strong" evidence according to the Jeffrey scale[9] and was used as the criterion for an outlier in our study. The direction of the outlier (*e.g.*, greater or lesser incidence of blood pressure [or Spo_2] noncompliance than the other anesthesiologists) was determined by the sign of the random effect term corresponding to the anesthesiologist. A negative δ_k indicated that the k^{th} anesthesiologist had a greater incidence of noncompliance relative to the other anesthesiologists.

Overall standard errors (SEs) of the incidences of blood pressure (or Spo_2) first checked 5 min or more after induction were calculated treating each of the periods as point estimates. This was because the same anesthesiologists were tested among periods.[12] The numerator and denominator for each period were used in the Freeman–Tukey transformation.[12] Student *t* distribution was used to calculate the CI and *P* value of the transformed values.[12] By using the harmonic mean number of anesthesiologists per period, incidences were back-transformed to the percentage incidences.[13] The SE was then calculated as the CI width divided by (2 × 1.96), the 1.96 being the inverse of the standard normal distribution. Coverage is accurate,[12] and the incidences and associated SEs are reported in the probability (percentage) scale. Along with the incidences of blood pressure and Spo_2 first checked ≥5 min after induction, 95% CIs were also provided.

Basic data analyses were performed by using the SAS software 9.3 (SAS Institute, Inc.). Classification tree analyses were performed by using SAS Enterprise Miner software 7.1. Plots were created using SigmaPlot version 12.5 (Systat Software, USA). Bayesian analysis were performed by using the WinBUGS 1.4.3 software (Imperial College and Medical Research Council, United Kingdom).[14]

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, and Rubin diagnostics plots,[15] 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.

Reporting Of Bayes Used in clinical STudies (ROBUST) guidelines was followed to report Bayesian analyses in this study.[16]

Results

Descriptive statistics are provided in table 1 for those variables used in one or more models for 68,220 general anesthetics. Descriptive statistics for those variables not used in

Table 1. Descriptive Statistics for Variables Used in One or Both of the Models (n = 68,220)

Variable	Mean ± SD	Median (Q25, Q75)
Blood pressure latency (minutes from first induction agent to first blood pressure)	−3.77 ± 87.62	−1.0 (−5.0, 1.0)
Spo_2 latency (minutes from first induction agent to first blood pressure)	−5.54 ± 78.25	−3.0 (−6.0, −1.0)
Case start latency (hours from the start of the surgical day to the time of induction)	3.74 ± 3.93	3.0 (0.37, 5.62)
Patient's age (yr)	41.70 ± 23.56	45.97 (22.0, 60.0)
American Society of Anesthesiologists physical status score	**%**	**N**
1	22.47	15,327
2	47.00	32,066
3	25.80	17,600
4	3.85	2,629
5	0.25	169
6	0.05	36
Missing	0.58	393

Spo_2 = pulse oximetry measured oxygen saturation.

models, based on lack of predictive value, are presented in tables 2 and 3, Supplemental Digital Content 1, http://links.lww.com/ALN/B145.

Blood Pressure Outcome

The unadjusted overall incidence of noncompliance for all 63,913 cases from all five periods for the blood pressure metric was 5.35% (95% CI, 5.17 to 5.52%). Summary results, incidences of noncompliance, and the number of anesthesiologists for each of the five 6-month periods are given in table 2 for blood pressure metric.

As summarized in the introduction, Ehrenfeld *et al.*'s[3] frequentist method uses the raw observed percentages for each anesthesiologist, without any covariate adjustment. Following their method, the upper 95% one-sided confidence limit (5.49%) was calculated for the overall (departmental) incidence of noncompliance. Based on this, 28.52% (18.68 to 38.36%) of the anesthesiologists in our department would be identified as noncompliant outliers (23 of 53, 14 of 56, 13 of 55, 16 of 59, and 14 of 57 in each of the five periods, respectively). Among the nine anesthesiologists each with at least 50% of their cases being pediatric (age <13 yr), all were frequentist outliers for at least four of five periods. Seven of these anesthesiologists were frequentist outliers for all five periods.

Applying the Bayesian analyses, 4.24% (3.18 to 5.30%) of anesthesiologists were outliers *without adjustment* for patients' ages. For example, anesthesiologist 3 (the anesthesiologist with the third greatest number of cases during the whole 2.5 yr) was detected as having a significantly greater incidence of blood pressure noncompliance than the other anesthesiologists in all five periods (table 2). The random chance of detecting this anesthesiologist as an outlier in all five periods was miniscule (1.82×10^{-9}). The fact that the same anesthesiologist was detected during different periods suggests the reliability of the Bayesian method.

Figure 4 shows an example of the process for a single 6-month period: January through June 2013. In this period, there were 14 anesthesiologists who had a significantly greater incidence of blood pressure noncompliance than the other anesthesiologists according to Ehrenfeld's frequentist model. In contrast, there were only two anesthesiologists who were outliers by the unadjusted Bayesian model. These two outlier anesthesiologists are represented by the rightmost solid green triangles (12.18 ± 1.65% and 18.96 ± 1.92% incidences of noncompliance in this period). The posterior probabilities for these two anesthesiologists each being an outlier were 53 and 99%, respectively.

When the Bayesian model was adjusted for the covariate of age, 2.44% (1.28 to 3.60%) of the anesthesiologists were detected as outliers. In the January 2013 through June 2013 period, the incidences of noncompliance of the two anesthesiologists with a significantly greater incidence of noncompliance than all other anesthesiologists (red squares in fig. 4) were 6.53 and 8.54%, respectively, for the covariate-adjusted model (table 2). The posterior probabilities for these two anesthesiologists each being an outlier were increased from 5% to 10% and to 73%, respectively. Although the incidences of outliers between the adjusted and unadjusted Bayesian models were not significantly different, covariate adjustment was important because these are different anesthesiologists (*i.e.*, the outliers without adjustment were not the same as those identified with adjustment) (table 2 and fig. 4).

Spo_2 Outcome

The overall incidence of noncompliance for the Spo_2 metric was 1.22% (95% CI, 1.14 to 1.30%) (table 3). If an anesthesiologist had an incidence of Spo_2 measurement within 5 min of induction that was greater than the upper 95% CI of overall incidence of noncompliance (1.29%), the anesthesiologist was judged as "noncompliant." Using Ehrenfeld

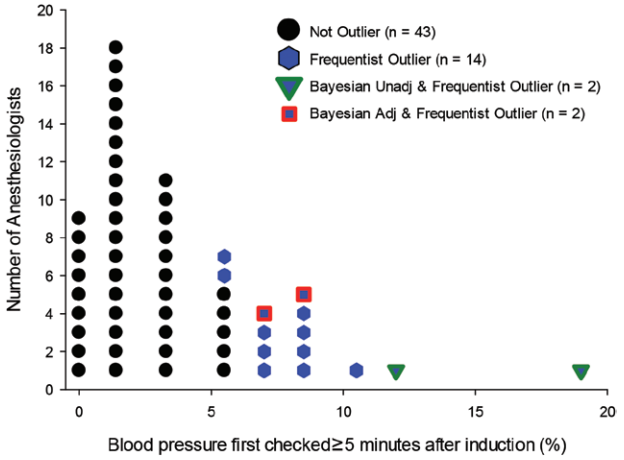

Fig. 4. Dotplot for January 2013 through June 2013 for the blood pressure metric. The *blue hexagons* represent the anesthesiologists with a significantly greater incidence of blood pressure first checked ≥5 min after induction than the other anesthesiologists based on the criteria of Ehrenfeld *et al*., without covariate and multiple comparison adjustment, as published by Ehrenfeld *et al*.[3] The *green triangle* shows the two anesthesiologists who were outliers when the Bayesian method was applied without covariate adjustment. The *red squares* show the two anesthesiologists who were outliers when the Bayesian method was applied with covariate adjustment. These four anesthesiologists are also frequentist outliers, which is why their *symbols* also include some *blue*. Note that, there are nine anesthesiologists with noncompliance rates of 0%, compliance rates of 100%, for the blood pressure metric in this period. Adj = adjusted; Unadj = unadjusted.

et al.'s method[3] for Spo_2, 40.20% (17.90 to 62.51%) of our anesthesiologists were detected as outliers (table 3). For example, the blue hexagons in figure 5 represent the 19 anesthesiologists who were frequentist outliers for the Spo_2 metric during the July 2011 through December 2011 period. In contrast, when the Bayesian model was used without adjusting for covariates, only one anesthesiologist (23) was deemed to be noncompliant and for just one period (0.11%). The posterior probability for this anesthesiologist being an outlier was increased from 5 to 35%. When the Bayesian model was adjusted for its covariates, none of the anesthesiologists had a significantly greater incidence of Spo_2 noncompliance than the others (table 3).

Sensitivity Analyses

For the results presented up to this point, the prior probability of *each anesthesiologist* having a significantly greater incidence of blood pressure or Spo_2 noncompliance than the other anesthesiologists was set to 5.0%. A sensitivity analysis was performed, with the prior probability of *at least one anesthesiologist* in the department during the studied 6-month period having a significantly greater incidence of blood pressure (or Spo_2) first checked 5 min or more after induction than the other anesthesiologists (tables 4 and 5, Supplemental Digital Content 1, http://links.lww.com/ALN/B145). For the blood pressure metric, fewer anesthesiologists were detected as outliers (pairwise differences 2.05% [−0.61 to 4.72%] for the unadjusted model and 2.10% [1.22 to 2.98%] for the adjusted model). For the Spo_2 metric, more were detected (0.11% [0.07 to 0.15%] for both unadjusted and adjusted models).

Discussion

Bayesian hierarchical outlier detection methods that take into account patient and practice characteristics provided more reliable and valid performance assessments for OPPE compared with those methods assessing raw incidence of compliance.

Comparison with the Frequentist Approach

The use of SAS Enterprise Miner enabled us to screen 135 potential covariates to learn what was important for our department. The methodology was an effective screening tool. However, the decision trees for our department are unlikely suitable for other departments. In other words, the "result" is the process, not the decision tree itself. We expected age to be a covariate for the blood pressure metric (*e.g.*, sevoflurane induction in a child with ASA physical status 1 for myringotomy tube placement followed by placement of the blood pressure cuff). However, the fact that no other variable was an important covariate was a surprise.

As we used decision trees, thresholds were not used for end-tidal concentrations (*e.g.*, >1% desflurane). The results were implausible clinically, and, from this, we recognized that there were residual subhypnotic amounts of agents present from the preceding case. Because we were analyzing 135 variables, 68,220 records, and hundreds of minutes of records, identifying this uncommon effect would otherwise have been challenging.

Figure 3 provides another example for why decision trees were useful, but our specific model should not be applied directly to other departments. At our hospital, transport monitor data were not automatically uploaded into the electronic medical record. The vital signs needed to be entered manually after transport had been completed. Our results made no differentiation based on when the vital signs were entered. This is a system-based informatics problem. Other hospitals are likely to have their own unique system-based challenges. Our generalizable result is the usefulness of the classification tree methodology to determine the department-specific covariates.

The two particular endpoints of our study used a threshold of 5 min. This choice is conservative, as it indicates that the patient did not have an Spo_2 checked until at least 5 min after recorded induction. This may also be a charting problem, but if so, it does reflect a responsibility of the supervising anesthesiologist.

Ehrenfeld *et al.*[3] used the same blood pressure metric, and, to take into account the patient's age, they excluded

Table 2. Summary Results for Each 6-month Period for the Blood Pressure Metric with the Prior Probability Set to 0.05

	January 2011 to June 2011	July 2011 to December 2011	January 2012 to June 2012	July 2012 to December 2012	January 2013 to June 2013
Number of anesthetics evaluated	11,799	13,392	13,408	13,571	11,743
Number of evaluated anesthesiologists supervising at least one anesthetic	53	56	55	59	57
Number of anesthetics per anesthesiologist, median (range)	207 (11–574)	220 (3–546)	212 (11–548)	201 (16–515)	181 (15–422)
Incidence of evaluated anesthetics with blood pressure noncompliance n (%)	761 (6.45%)	728 (5.4%)	717 (5.3%)	666 (4.9%)	545 (4.6%)
Anesthesiologists identified as performance outliers					
Frequentist	n = 23	n = 14	n = 13	n = 16	n = 14
Bayesian unadjusted (anesthesiologist identifier)	n = 2 (3, 49)	n = 3 (3, 20, and 38)	n = 2 (3 and 38)	n = 3 (1, 3, and 25)	n = 2 (1 and 3)
Bayesian adjusted (anesthesiologist identifier)	n = 1 (3)	n = 2 (3 and 38)	n = 1 (38)	n = 1 (25)	n = 2 (10 and 52)

The unadjusted overall incidence of noncompliance for all 63,913 patients from all five periods for the blood pressure metric was 5.35% (95% CI, 5.17–5.52%) (3,417 of 63,913). Anesthesiologists were labeled according to their number of anesthetics during the whole 2.5-yr period. For example, anesthesiologist 1 performed the most number of anesthetics and anesthesiologist 2 is the second most in anesthetics. The adjusted model includes the patient's age. In this table, the prior probability of each anesthesiologist having a significantly greater incidence of blood pressure noncompliance than the other anesthesiologists was set to 0.05. For example, the last column represents results from January 2013 through June 2013 for the blood pressure outcome. In that period, 57 anesthesiologists performed 11,743 general anesthetics. The number of anesthetics per anesthesiologist ranged between 15 and 422. The incidence of noncompliance for checking blood pressure within 5 min after induction in this period was 4.6%. Cochran–Armitage test of trend indicates reducing raw incidences of noncompliance over time ($P < 0.0001$).

pediatric cases. Entirely excluding a class of patients is not covariate adjustment. Our results show that, for a metric such as Spo_2, excluding all cases coming from an ICU and all the first case starts would result in inaccurate identification of the anesthesiologists not meeting the desired performing standard. In addition, Ehrenfeld *et al.*'s[3] frequentist approach excluded anesthesiologists performing few anesthetics (*e.g.*, 35% excluded by Ehrenfeld *et al.*, see Materials and Methods). Because OPPE by definition is to be applied to all anesthesiologists in a department, we need methods for risk adjustment that function without excluding classes of patients and/or anesthesiologists (see first section of Materials and Methods).

We used a random effects model to represent heterogeneity among anesthesiologists. Because of the preassumption that anesthesiologists are performing similar to each other in the Bayesian hierarchical model, those anesthesiologists with much greater or lesser incidence of noncompliance compared with the other anesthesiologists were shrunk toward the overall mean. This implies that, if an anesthesiologist was detected as an outlier based on the Bayesian analysis, the anesthesiologist's performance should truly be outlying. Due to the shrinkage toward the overall mean, it is hard to detect an anesthesiologist with few cases as an outlier with the Bayesian method. However, with the Ehrenfeld *et al.*'s[3] approach, the anesthesiologists with few cases are not even analyzed.

Ehrenfeld *et al.*[3] used frequentist outlier detection without covariate adjustment. We used the decision tree for the blood pressure outcome (fig. 1) to show that adding covariate adjustment to the frequentist analysis would not substantively address the weaknesses of the approach. We started by deleting the group of patients with the greatest incidence of the blood pressure not being checked within 5 min after induction (*i.e.*, children <7 yr and 3 months). When this age group was deleted, the incidence of noncompliance among the remaining cases was 3.14% (1,876 of 59,689). Applying the frequentist analysis, anesthesiologists are detected as outliers if they have an incidence greater than the 95% upper confidence limit for the overall incidence. This condition was satisfied by 43.86% (25 of 57) of our department's anesthesiologists during the January through June 2013 period. This result demonstrates that using frequentist analyses with covariate adjustment is not a useful tool for assessing anesthesiologists' performance as outlier *versus* not outlier.

Mathematically, it may seem a substantial leap in complexity to go from Ehrenfeld *et al.*'s[3] simple method (based on observed incidence of noncompliance with no covariate adjustment and no multiple comparisons adjustment) to a Bayesian logistic regression model (with adjusted covariates chosen using data mining). However, there are no methods that mathematically are "in between," and what methods are available are not simpler to perform. In sequence, covariates need to be chosen with interactions while incorporating

Table 3. Summary Results for Each 6-month Period for Spo_2 with the Prior Probability Set to 0.05

	January 2011 to June 2011	July 2011 to December 2011	January 2012 to June 2012	July 2012 to December 2012	January 2013 to June 2013
Number of anesthetics evaluated	11,799	13,392	13,408	13,571	11,743
Number of evaluated anesthesiologists supervising at least one anesthetic	53	56	55	59	57
Number of anesthetics per Anesthesiologist, median (range)	207 (11–574)	220 (3–546)	212 (11–548)	201 (16–515)	181 (15–422)
Incidence of evaluated anesthetics with Spo_2 noncompliance n (%)	229 (1.94%)	156 (1.16%)	153 (1.14%)	141 (1.04%)	100 (0.85%)
Anesthesiologists identified as performance outliers					
Frequentist	n = 37	n = 19	n = 25	n = 18	n = 13
Bayesian unadjusted (anesthesiologist identifier)	n = 0	n = 1 (23)	n = 0	n = 0	n = 0
Bayesian adjusted	n = 0	n = 0	n = 0	n = 0	n = 0

The overall incidence of noncompliance for the Spo_2 metric was 1.22% (95% CI, 1.14–1.30%) (779 of 63,913). Anesthesiologists were labeled according to their number of anesthetics during the whole 2.5-yr period. For example, anesthesiologist 1 performed the most number of anesthetics, and anesthesiologist 2 is the second most anesthetics performed. The adjusted model includes covariates. ASA is 1 when the ASA physical status score is ≥4 and 0 otherwise. Start of the day is a binary variable indicating if the time from the start of the surgical day to induction was ≤5 *vs.* >5 min. The "From ICU" variable reports if the preceding location before the patient was in an OR was an ICU. For interpretation, see the legend of table 2. Cochran–Armitage test of trend indicates reducing raw incidences of noncompliance over time ($P < 0.0001$).

ASA = American Society of Anesthesiologists; ICU = intensive care unit; OR = operating room; Spo_2 = pulse oximetry measured oxygen saturation.

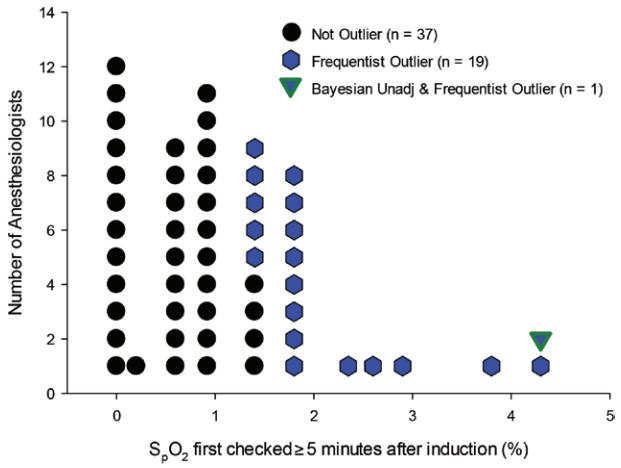

Fig. 5. Dotplot for July 2011 through December 2011 for the pulse oximetry measured oxygen saturation (Spo_2) metric. The *blue hexagons* represent the anesthesiologists with a significantly greater incidence of Spo_2 first checked ≥5 min after induction than the other anesthesiologists based on the criteria of Ehrenfeld *et al.*,[3] without covariate and multiple comparison adjustment. The *green triangle* shows the single anesthesiologist who was an outlier when the Bayesian method was applied without covariate adjustment. This anesthesiologist is also a frequentist outlier, which is why the *symbol* also includes some *blue*. None of the anesthesiologists was detected as having a significantly greater incidence of Spo_2 first checked ≥5 min after induction than the other anesthesiologists when the Bayesian method was applied with covariate adjustment. Unadj = unadjusted.

incidences of the covariates (fig. 3). That is a regression tree. The logical alternative approach to use would be neural networks. They are (most assuredly) not simpler to use. Next, we used a risk-adjusted Bayesian model. A frequentist logistic regression model could be used, instead, but that would not be the logistic regression of basic statistical packages because such packages do not include the random anesthesiologist effect.

Sensitivity Analyses

The incidence of overall noncompliance for the Spo_2 metric (1.22 ± 0.04%) was much lower compared with the blood pressure metric (5.35 ± 0.09%). Because compliance rates for each anesthesiologist were high and closer to each other for the Spo_2 metric, the variances of the random anesthesiologist effects for the Spo_2 metric were less compared with the variance for the of the random anesthesiologist effects for the blood pressure metric for both adjusted and unadjusted model. As explained in the Materials and Methods section, for each anesthesiologist, setting the individual prior probability to 0.05 is a greater prior probability compared with setting the prior probability for overall departmental probabilities. Having a smaller variance for the Spo_2 metric in the data and having a smaller probability with the overall departmental prior distribution together led to having a smaller variance for the overall departmental prior distribution situation compared with the individual probability. Therefore, slightly more (0.11 [0.07 to 0.15]) individuals were detected as outliers for the Spo_2 metric when the overall departmental probabilities were used.

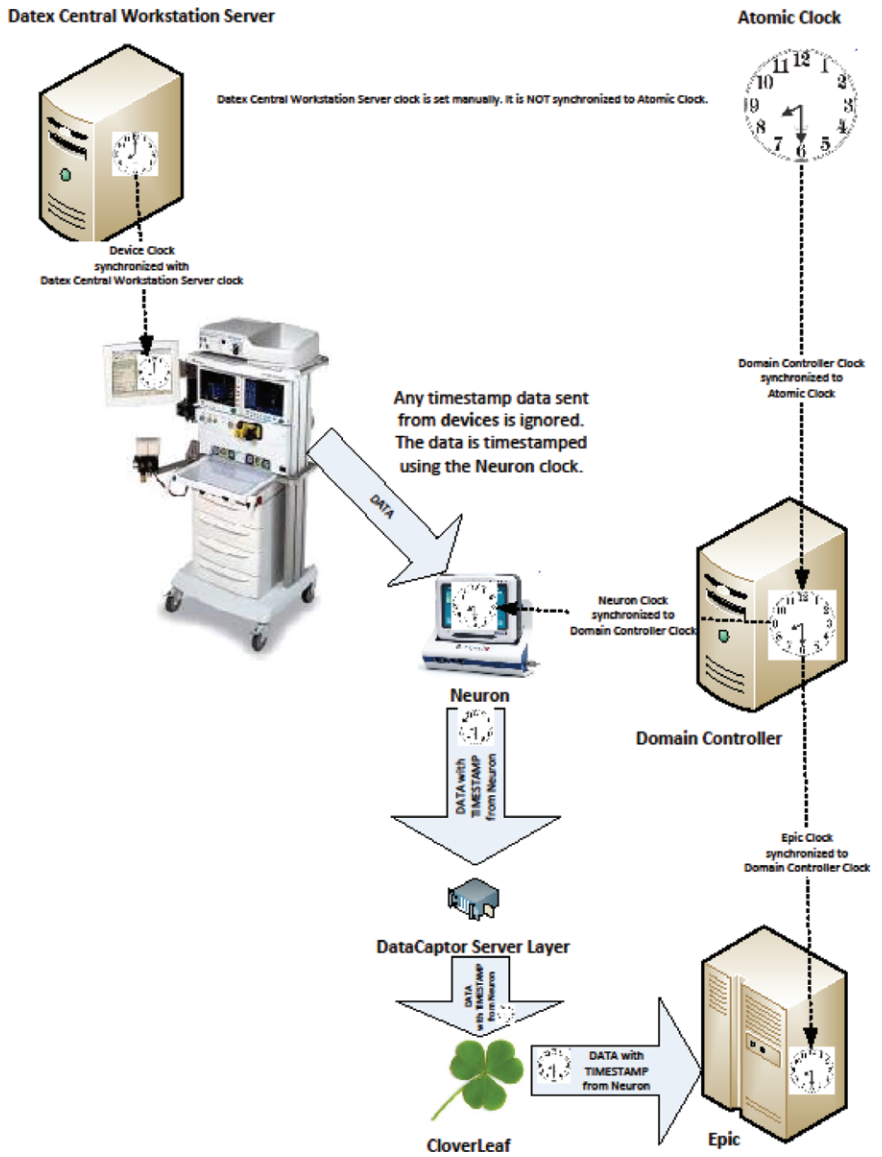

Fig. 6. Schematic illustration of the transmission of the data and the synchronization of the time stamps. Epic = Epic anesthesia information management system (Epic Systems, Inc., USA).

Limitations

Our study has limitations. First, the outcomes we have chosen for the demonstration of our Bayesian methods may not reflect the actual performance levels of anesthesiologists. For example, an anesthesiologist may check blood pressure from a transport monitor without reporting it in the electronic medical records within 5 min after induction. This is a systems-based problem. Although charting is part of medical care, it is not that the blood pressure was not being measured. Second, we analyzed the anesthesiologist, but anesthesia at our hospital is a team with anesthesiology resident(s) and/or Certified Registered Nurse Anesthetists. Third, we excluded those anesthesiologists (rotating fellows and locum workers) who did not work for the department for each 6-month period that was studied. Our methods could have included these groups of anesthesia providers. As explained in the "Comparison with the Frequentist Approach" section, due to the shrinkage toward the overall mean, the method will appropriately be unlikely to detect an anesthesiologist with a too small sample size as a Bayesian outlier. Therefore, including those anesthesiologists who worked for the department for less than 6 months would not change the results of our Bayesian analyses for our department but would have made the proportional effect of anesthesiologists as outliers artificially less.

Another fundamental issue is the lack of any "definitive standard" for performance. OPPE only requires a comparison of providers within a department. Therefore, our approach compares providers with their peers—not *versus* some external standard. It is possible that the overall departmental incidences are inappropriate (*e.g.*, the greater incidence of unreported Spo_2 before induction of anesthesia among patients from ICUs as shown in fig. 3). Consequently, we

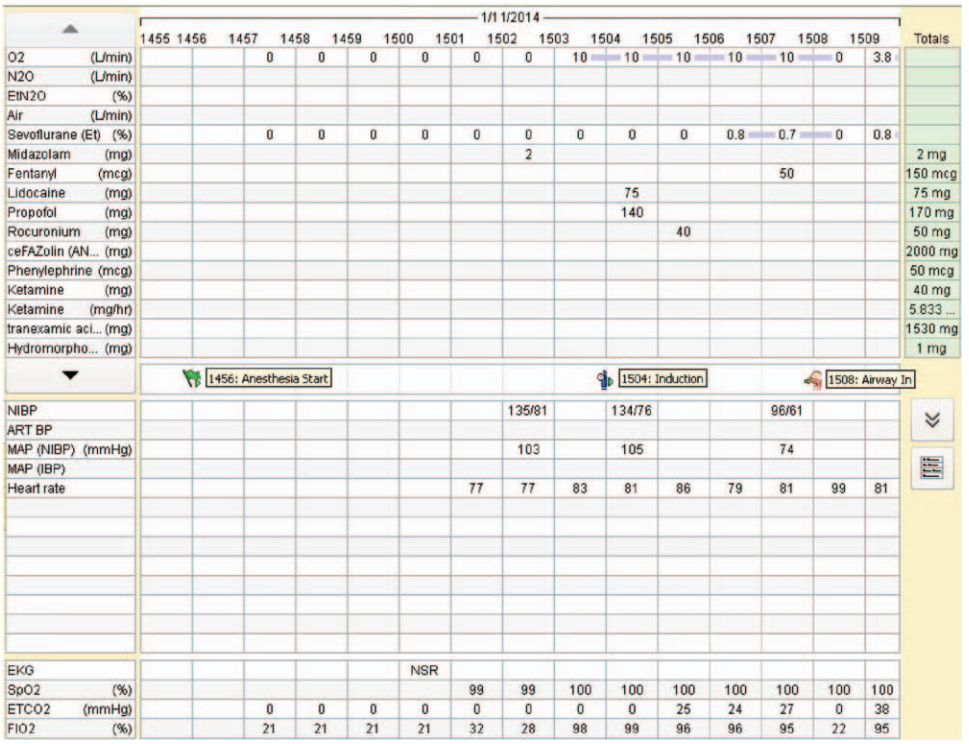

Fig. 7. Illustration of Epic (Epic Systems Corporation, USA) records for using thresholds for end-tidal concentrations. ART = arterial; EKG = electrocardiogram; IBP = invasive blood pressure; MAP = mean arterial pressure; NIBP = noninvasive blood pressure; Spo_2 = pulse oximetry measured oxygen saturation. Reproduced, with permission, from © 2015 Epic Systems Corporation.

cannot claim that the "lower" incidence of outliers seen with our approach is "correct," whereas the unadjusted frequentist approach is "incorrect." Nevertheless, the suggestion (based on unadjusted frequentist methods) that nearly half of all of our anesthesiologists are "outliers" at some point in time seems unlikely. Neither our approach nor the frequentist method can detect "global" noncompliance problems (*e.g.,* everyone within a department failing to measure blood pressure before induction). However, any such global standard (and compliance) would still need to be established *via* some kind of covariate adjustment process, similar (we believe) to that described in our study.

Conclusions

Given that the use of a Bayesian hierarchical multivariate methodology takes into account patient and practice characteristics, it is more representative of differences in case numbers and case mix of the anesthesiologists compared with a nonhierarchical frequentist approach. Therefore, Bayesian hierarchical methods may be a preferable method for mandated monitoring of the performance of anesthesiologists instead of those methods assessing the raw incidence of compliance.

Acknowledgments

The authors thank David Griffiths, B.S., and Gregory Hopson, B.A., M.I.S., both from the Department of Anesthesia, University of Iowa, Iowa City, Iowa, for their assistance with extracting the data from Epic anesthesia information management system 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 52242. 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 Informatics and the Bayesian Model

Anesthesia Medical System Time Stamps

Figure 6 shows the schematic illustration of the transmission of the data and the synchronization of the time stamps. Note that, there is no rounding on the time stamps of the data from the monitors. For example, the time stamp of the data 7:59:59 AM seen from the monitor is recorded as 7:59 AM. Similarly, the time stamp data of 8:00:01 AM is 8:00 AM. Because there is no rounding on the time stamps of the data,

0 to 2 min of difference on the time stamp of the data and the monitor is expected. In other words, we would get artifact from the engineering description of NEURON using the integer portion of minutes if we want to monitor if the anesthesiologist checked blood pressure within 2 min after induction. The transmission time is instantaneous when the device sends the data. However, the interval at which the data are sent varies by devices and can vary within the device.

Examples of Calculating the Total Dose of Propofol

Total propofol dose after the anesthesia start time during the first 5 min of initial propofol administration was calculated. Both bolus dose and infusion dose were taken into account. Assume the patient is 70 kg. Examples:

1. At 11:04:00 AM, an infusion of propofol is started at a rate of 150 μg kg^{-1} min^{-1}. The anesthesia start time is listed as 11:05:00 AM. The total propofol dose used in calculations from 11:05:00 AM to 11:09:59 AM is 42 mg, where 42 mg = (150 μg kg^{-1} min^{-1}) × (70 kg) × (4 min)/1,000.
2. Anesthesia start time is 11:20:00 AM, 50 mg propofol is administered at 11:32:00 AM, and an infusion rate of 75 μg kg^{-1} min^{-1} propofol started at 11:36:00 AM. The 5-min window is 11:32:00 to 11:36:59 AM. Therefore, propofol infusion only from the last minute is used. The total propofol dose = 50 + 70 × 75 × 1/1,000 = 55.25 mg.
3. Anesthesia start time is 11:20:00 AM, 50 mg propofol is administered at 11:32:00 AM, and an infusion of 75 μg kg^{-1} min^{-1} propofol started at 11:40:00 AM. The 5-min window is 11:32:00 to 11:36:59 AM. Therefore, propofol infusion is not added to the calculation. The total propofol dose is 50 mg.

Using Thresholds for End-tidal Concentrations

Consider the patient presented in figure 7 with anesthesia start time of 2:56 PM and induction time of 3:04 PM. The first arterial blood pressure or first noninvasive blood pressure after the anesthesia start time was at 3:02 PM for this patient.

The time of first agent (propofol, rocuronium, etomidate, sevoflurane, and desflurane) was calculated. When thresholds were not used for three volatile anesthetics, the time of first agent was reported as 2:57 PM, indicating the first sevoflurane time. The blood pressure latency minute is the time from the first agent (2:57 PM) to the first blood pressure recording (3:02 PM) and becomes 5 min. In this example, the anesthesia provider's blood pressure outcome would be considered noncompliant for this patient. When thresholds were used for the volatile agents, the first reading of sevoflurane is at 3:06 PM. The first agent time is 3:04 PM for propofol, and blood pressure latency minute is −2 min. Therefore, the blood pressure outcome of the anesthesia provider for this patient becomes compliant.

The Bayesian Model

Let n_k denote the number of anesthetics with induction by anesthesiologist k ($k = 1, \ldots, K$) where K is the current number of anesthesiologists in the department in a given 6-month period. $y_{ik} = 1$ denotes compliance with a metric (*e.g.*, pulse oximetry measured oxygen saturation [Spo_2] checked before induction) for anesthetic i ($i = 1, \ldots, n_k$) for anesthesiologist k and $y_{ik} = 0$, otherwise. Assuming each anesthesiologist's compliance is independent of another anesthesiologist, y_{ik} are Bernoulli random variables, and the probability of compliance can be denoted by p_{ik}. In other words:

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

The logit link is used to normalize the compliance rates. The log odds of a compliance for anesthetic i with anesthesiologist k is denoted by $\theta_{ik} = \text{logit}(p_{ik}) = \log(p_{ik}/(1\text{-}p_{ik}))$.

θ_{ik} can be written as a function of patient and surgical characteristics. For example, the final model for the Spo_2 metric with the significant covariates can be written as follows:

$$\theta_{ik} = \mu + \beta_1\ \text{ASA} + \beta_2\ \text{fromICU} + \beta_3\ \text{HoursFromStart} + \delta_k$$

where μ is the intercept in the logit scale, β_1 to β_3 are coefficients for the independent covariates, and δ_k is the random anesthesiologist effect. It should be noted that these parameters were defined on the logit scale. ASA is 1 when the ASA physical status score is 4 or greater and 0 (zero) otherwise. "from ICU" is 1 when the patient comes from ICU and 0 otherwise. Similarly, "HoursFromStart" is 1 when the minutes from the start of the day were greater than 5 min and 0 (zero) otherwise.

Under the exchangeability[9] assumption, anesthesiologists are considered to be sampled from a common distribution. This is assumed to be a normal distribution with a mean of 0 and an SD of σ. In mathematical notation, this can be written as follows: $\delta_k \sim \text{normal}(0, \sigma^2)$. This reflects the similarity and differences between anesthesiologists.

Although only anesthesiologists with greater incidences of checking blood pressure or Spo_2 before induction are presented in this study, with the proposed method, it is possible to detect also anesthesiologists with incidences that are less than average. The same steps can be followed. To identify whether an anesthesiologist has an incidence greater or less than average, the sign of delta is assessed. The same Bayes factor cutoff (Bayes factor <0.1) can be used for the strong evidence of having a lesser incidence than average.

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.[10] Weakly informative prior distributions were used for the overall mean, μ, and the coefficients for the fixed effects, β_1 to β_3. Namely, the prior distribution for the overall mean was assumed to have a normal distribution with a mean of 0 and an SD of 2; $\mu \sim \text{Normal}(0, 2^2)$. 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)).

Prior distributions for binary covariates (ASA physical status, whether the patient was coming from an ICU, and

hours from the start of the day, ≤5 *vs.* >5 min) were also assumed to be normal distribution with a mean of 0 and an SD of 2; $\beta_i \sim \text{Normal}(0, 2^2)$.

As explained on the Materials and Methods section, two times the square-root transformation provided closest to a linear relation with the logit probability of blood pressure first checked 5 min or more after induction. Because age with two times the square-root transformation was on the continuous scale and has a wider scale, the SD of this prior normal distribution was assumed to be 1: $\beta_i \sim \text{Normal}(0, 1^2)$.

When the prior distribution includes the available prior knowledge, it is called an "informative" prior distribution. An informative inverse-gamma prior distribution is used for the between-anesthesiologist variance σ^2: $\sigma^2 \sim$ Inv-Gamma(α,β), with a mean of 0.125 and an SD of 0.05.

Individual Prior Probability

Chaloner and Brant[7] defined an 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 distribution of ε_is are independent and normally distributed with a 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 is an outlier:

$$\Pr(|\varepsilon_i| > m) = \Pr(\varepsilon_i > m) + \Pr(\varepsilon_i < -m) = 2 * \Phi(-m)$$

where Φ (z) is the standard normal distribution function.

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 0.05. In this circumstance, m becomes 1.96.

Overall Prior Probability

As a sensitivity analysis, the prior probability of an anesthesiologist being an outlier can be calculated from departmental norms. As explained in the previous section, the prior probability that the k^{th} anesthesiologist is an outlier can be written as: $\Pr(|\varepsilon_i| > m) = \Pr(\varepsilon_i > m) + \Pr(\varepsilon_i < -m) = 2 \times \Phi(-m)$.

The prior probability that the k^{th} anesthesiologist is NOT an outlier equals: $1–2 \times \Phi(-m)$.

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:

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

The prior probability of not detecting any anesthesiologist in the department, during the studied 6-month period, having an outlier incidence of blood pressure (or Spo_2) first checked 5 min or more after intubation should be high and is set to 95%. In other words:

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

This corresponds to the probability of "at least one anesthesiologist in the department during the studied 6-month period having a significantly greater incidence of blood pressure (or Spo_2) first checked 5 min or more after induction than the other anesthesiologists" being equal to 0.05.

The prior probability of one specific anesthesiologist k being an outlier when there are K anesthesiologists is $2\Phi(-m)$, where:

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

For example, in January 2013 through June 2013, there were 57 anesthesiologists, K = 57. Thus, m = 3.320, and the prior probability of being outlier for a specific anesthesiologist is 0.0009.

Adjusted WinBUGS Model for Blood Pressure Outcome - Individual Probability

```
Model
{
A   for (i in 1:11743){
B           goodoutBP[i] ~ dbern(p[i])
C           logit(p[i]) <- theta[i]
D           theta[i] <- mu + beta1*age[i] + delta[anes[i]]
          }
E   for(k in 1:57){
F       Post.delta.1.96[k] <- step(delta[k] – 1.96*sigma.e) +
        step(-delta[k] – 1.96*sigma.e)
        #> m <- qnorm(0.975) = 1.96
              }
G   Prob.sum <- sum(Post.delta.3.32[])
H   Prob.any.g1.96 <- step(Prob.sum -1)

    for(j in 1:57){
I     delta[j] ~ dnorm(0, prec.delta)
    }
J   mu~ dnorm(0, 0.25) # sd = 2
K   beta1~ dnorm(0, 1) # sd = 1
L   prec.delta ~ dgamma(9, 1)
M   sd.delta <- 1/sqrt(prec.delta)
}
```

At **A**, i refers to the i^{th} of 11,743 anesthetics performed in January 2013 through June 2013.

At **B**, goodoutBP[i] is a binary variable denoted by 1 if the anesthetic i is compliant for the blood pressure metric and 0 otherwise. Compliance rate has a Bernoulli distribution with the compliance probability of p[i].

At **C**, the logit transformation is applied to the compliance rate, similar to the logistic regression model.

At **D**, the compliance probability is written as a function of overall intercept (μ), patients' age, and the random anesthesiologist effect (delta[anes[i]]).

At **E**, *k* stands for the k^{th} of 57 anesthesiologists who is assessed for performance during the January 2013 to June 2013 period.

At **F**, the posterior probability of being outlier for anesthesiologist *k* (Post.delta.1.96[k]) is calculated. The step function is used to calculate how many times the random anesthesiologist effect is more extreme than $m = 1.96$. Step(e) returns 1 if $e \geq 0$ and 0 (zero) otherwise.

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

At **H**, the probability of at least one provider being an outlier (Prob.any.g1.96) is calculated.

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

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

At **K**, a prior distribution is defined for the slope of patient's age (Box-Cox transformed). This is a normal distribution with a mean of 0 and an SD of 1 and, therefore, a weak informative prior distribution.

At **L**, the prior distribution for the precision of the random provider effect is defined as a gamma distribution with parameters 9 and 1. The mean of this gamma distribution is 9 and its SD is 3.

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

References

1. Haller G, Stoelwinder J, Myles PS, McNeil J: Quality and safety indicators in anesthesia: A systematic review. Anesthesiology 2009; 110:1158–75
2. Ohlssen DI, Sharples LD, Spiegelhalter DJ: Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. Stat Med 2007; 26:2088–112
3. 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
4. Bayman EO, Chaloner K, Cowles MK: Detecting qualitative interaction: A Bayesian approach. Stat Med 2010; 29:455–63
5. Ehrenfeld JM, Epstein RH, Bader S, Kheterpal S, Sandberg WS: Automatic notifications mediated by anesthesia information management systems reduce the frequency of prolonged gaps in blood pressure documentation. Anesth Analg 2011; 113:356–63
6. de Ville B, Neville P: Decision Trees for Analytics Using SAS Enterprise Miner. Cary, North Carolina, SAS Institute, 2013
7. Chaloner K, Brant R: A Bayesian approach to outlier detection and residual analysis. Biometrika 1988; 75:651–9
8. 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
9. Spiegelhalter DJ, Abrams KR, Myles JP: Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, John Wiley & Sons, 2004, pp 55–92
10. Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Anal 2006; 1: 1–19
11. Kass RE, Raftery AE: Bayes factors. J Am Stat Assoc 1995; 90: 773–95
12. Dexter F, Marcon E, Epstein RH, Ledolter J: Validation of statistical methods to compare cancellation rates on the day of surgery. Anesth Analg 2005; 101:465–73
13. Miller JJ: The inverse of the Freeman–Tukey double arcsine transformation. The American Statistician 1978; 32:138
14. Lunn DJ, Thomas A, Best N, Spiegelhalter D: WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. Stat Comput 2000; 10: 325–37
15. Brooks SP, Gelman A: General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998; 7: 434–55
16. 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