

# Selection of Reliable and Valid Surgeon Performance Measures

**William Thomas Cecil**

Independent Consultant, Knoxville, Tennessee, the United States

**Email address:**

bcecill@chartertn.net

**To cite this article:**

William Thomas Cecil. Selection of Reliable and Valid Surgeon Performance Measures. *American Journal of Management Science and Engineering*. Vol. 5, No. 5, 2020, pp. 62-69. doi: 10.11648/j.ajmse.20200505.12

**Received:** October 26, 2020; **Accepted:** November 3, 2020; **Published:** November 11, 2020

---

**Abstract:** *Objective-* To identify measures of surgeon performance that are valid, reliable, and capable of classifying the risk of surgeon performance. *Data Sources-* A surgical quality improvement program, dataset unique to selected hospitals and surgeons containing abstracted surgical case records. *Study Design-* Six criteria were employed to assess the validity of 24 candidate measures of surgeon performance: 1) the presence of a surgeon random intercept; 2) a surgeon signal that is greater than zero; 3) surgeon majority control; 4) reliability of the surgeon random intercept of at least 0.7; 5) the capacity to identify both low- and high-risk surgeons and 6) the presence of a learning/improvement effect. *Data collection/Extraction methods-* Surgical case review nurses abstracted cases for each surgeon using a structured sampling and abstraction methodology. *Principal findings-* Comparing outcomes requires risk adjustment and the use of the "true score" approach but is limited by case volume constraints and a confounding factor, i.e., the hospital, if used to judge surgeons' performance. Assessing surgeon performance requires a measure of the surgeon's effects on the consequences (postoperative occurrences) of surgical procedures, i.e., the surgeon-specific random intercept, which is a product of a multilevel risk adjustment model. *Conclusion-* Morbidities and mortality lack the characteristics necessary to be used as measures of surgeon performance. However, the process (task-time) measures LOS and OT both have high event rates, high reliability, and are capable of classifying surgeon risk.

**Keywords:** Multilevel Mixed-Effects Modeling, Risk Adjustment for Clinical Outcomes, Reliability, Validity

---

## 1. Introduction

Surgeon performance measurements are potentially helpful for quality improvement [1], consumer decision support [2], and surgeon management [3, 4]. Models of the surgeon role in modern multidisciplinary care include the "captain of the ship" and "member of the team" models [5, 6]. According to the "captain of the ship" model, the surgeon assumes responsibility for patient and intervention selection. In contrast, in the "member of the team" model, decisions are made by the team. In a published statement regarding physician-led team-based surgical care, the American College of Surgeons (ACS) endorsed the team approach: "Optimal care is best provided by a coordinated multidisciplinary team recognizing each member's expertise. Coordinated surgical care provides the best outcomes, lowers costs, and increases patient satisfaction" (Statement on Physician-Led Team-Based Surgical Care) [7].

Recent studies investigating surgeon performance have

focused on establishing the feasibility of evaluating surgeon performance and reliability using discrete measures. However, few studies have focused on identifying surgeon performance [8, 9], and no studies have compared surgeon-related risk and demographic, preoperative condition, and surgical procedure-related risk.

Iezzoni proposed that the purpose of risk adjustment is to obtain "meaningful comparisons within the health care system that generally require risk adjustment—accounting for patient-associated factors before comparing outcomes across different patients, treatments, providers, health plans or populations [10]." The true risk score is the sum of the fixed and random effects identified by a multilevel mixed-effects risk model. The fixed effects consist of patient demographic factors, indicators of the presence or absence of patient preoperative conditions thought to impact the prevalence of postoperative complications and case-mix factors that reflect surgical procedure risk. In a three-level risk model, the random effects are estimated for the risk

added by the surgeons and hospitals. The "true score" used to assess surgeon performance compares the sum of the fixed and random effects to the sum of the fixed effects. It is expressed as the relative risk or odds ratio of a postoperative complication. In a three-level system (patient, surgeon, and hospital), the "true score" is a patient-level measure.

A measure's validity is affected by adequate observations, the performance measure's prevalence, and the sample size. Adams offered the following list of validity determinants for physician measures: 1) the level of physicians' control over the measure, 2) proper adjustment of case-mix differences among physicians, 3) whether another level in the system partially controls the measure, and 4) whether the measure is correlated with other established quality measures [11]. The level of surgeons' control over the candidate measures of surgeon performance has not previously been assessed.

The aim is to understand the impact of the constraints of validity, reliability, and model specification on the selection of surgeon performance measures.

## 2. Methods

### 2.1. Evaluation Framework

The following criteria were applied to determine the suitability of the candidate measure as a surgeon performance measure: 1) the presence of a surgeon random intercept; 2) a surgeon signal that is greater than zero; 3) surgeon majority control; 4) reliability of the surgeon random intercept of at least 0.7; 5) the capacity to identify both low- and high-risk surgeons and 6) the presence of a learning/improvement effect.

### 2.2. Risk Model

Twelve months of abstracted data included in the dataset for this study, with 29,267 surgical cases, 644 surgeons, and 23 hospitals, was used to evaluate 24 candidate measures of surgeon performance. The candidate measures included the following postoperative occurrences: mortality, acute renal failure (ARF), bleeding/transfusion (BT), cardiac arrest requiring CPR (CPR), deep incisional surgical site infections (DSSI), deep venous thrombosis (DVT), myocardial infarction (MI), ventilator use for more than 48 hours (ONVENT), organ/space SSI (OSSI), pneumonia (PNA), progressive renal insufficiency (PRI), pulmonary embolism (PE), sepsis, septic shock (SHOCK), stroke/cerebrovascular accident (CVA), superficial SSI (SSSI), unplanned intubation (UI), urinary tract infection (UTI), wound disruption (WD), patients with morbidity (PTSWMB), readmission (READ), return to the operating room (ROR), operative time (OT) and length of stay (LOS). In total, 15,366 inpatient cases were used to risk-adjust the LOS. Only one procedure was performed in 19,412 cases, which were used to risk-adjust the OT. Cases with multiple procedures are likely to confound the OT risk adjustment and were thus excluded from the OT analysis. The dataset was generated by surgical case reviewers based on sampled cases reported by surgeons.

Multilevel mixed-effects models appropriate to the type of postoperative occurrence were used for risk adjustment. A logistic model was used for binary occurrences (all except for the LOS and OT). A negative binomial model was used for the LOS (in days). A linear model was used for OT (in hours rounded to the nearest 0.01). Random intercepts were included at the second and third levels of the models, i.e., surgeons and hospitals, respectively. A three-level model was used to estimate the patient risk score because it reflects the patient, surgeon, and hospital system in which the surgery and postoperative occurrences occur.

Standard demographic, preoperative risk factors, and procedure identifiers were included as covariates. Variables for patient age, gender, body mass index, number of procedures per case, procedure groups, and a Current Procedural Terminology (CPT) code-based measure of postoperative occurrence risk were employed. In this study, the grouping method was based on 47 categories of CPT codes representing different procedures, such as hernia repair, colectomy, and vascular bypass/repair. The CPT code-based measure of each postoperative occurrence risk was estimated by constructing multilevel mixed-effects models; a random intercept was created for each CPT code used in previous periods as an independent variable in the risk model.

### 2.3. Model Specification

Each risk model was tested to determine whether a multilevel model is required using the likelihood ratio test to compare the model to the standard regression. A significant likelihood ratio test indicated that the multilevel model had an improved performance over the standard regression and that at least one of the additional levels was helpful. The variance components were analyzed to test the hypothesis that the between-surgeon within-hospital variance is zero for each candidate measure of surgeon performance using the likelihood ratio test to compare the full three-level model to an otherwise identical model in which the between-surgeon variance was set to zero by removing the surgeon random effect. If the hypothesis is correct, the surgeon signal and surgeon random intercept are not significant, and the surgeon's performance cannot be assessed.

### 2.4. Measuring and Classifying Surgeon Performance

According to each candidate measure, the surgeon's performance was compared using the Bayesian posterior mean (random intercept) of each surgeon and the 95% prediction interval. The values were assigned to the random intercept using empirical Bayes predictions based on the following obtained estimates: covariate coefficients ( $\beta$ ), between-surgeon variance ( $\psi^2$ ), and within-surgeon variance ( $\theta^2$ ). According to Bayes' theorem for linear models, the posterior distribution (posterior means/surgeon random intercepts) is proportional to the prior distribution multiplied by the likelihood of the responses.

The prior is a vector of shrinkage factors, and the likelihood is the surgeon's specific mean total residual. Surgeons with random intercepts and 95% prediction

intervals above zero had significantly larger intercepts than the all-case-averaged intercept, and the postoperative occurrence risk was higher. In comparison, surgeons who had 95% prediction intervals below zero had intercepts significantly smaller than the all-case-averaged intercept, and the risk of a postoperative occurrence was lower.

### 2.5. Reliability

The surgeon performance measure's reliability was calculated as the ratio of the between-surgeon variance to the sum of the between- and within-surgeon variance. The variance of the surgeon random intercepts, which is reported as a random surgeon effect in the multilevel mixed-effects model, is the between-surgeon variance. The within-surgeon variance is the squared standard error of the measurement (random intercept), which is reported as the standard error of the empirical Bayes estimator of the random effects. A reliability score of 0.7 was used as the required threshold for identifying a surgeon as having a high or low risk for any postoperative occurrence [12].

### 2.6. Measure Validity

To test Adam's first and third validity criteria, i.e., the physician control level, the between-surgeon and hospital variances (signals) estimated by the risk models were compared to identify the system's level with majority control over the candidate measure. A postoperative risk model in which the surgeon has a larger signal than the hospital, suggests that the surgeon has majority control and can be used as a surgeon performance measure if the other criteria are met. Spearman's rank correlation coefficients were used to assess the candidate measures' correlations and test Adam's fourth criterion.

### 2.7. Sensitivity Analysis of the Reliability Assessment

The surgeon reliability assessments using a two-level model, in which the random intercept is estimated only for the surgeons, was compared to the current three-level model, in which both the surgeon and hospital random intercepts are generated. If surgeon only models had higher reliabilities it would confirm the need to use a hospital random intercept.

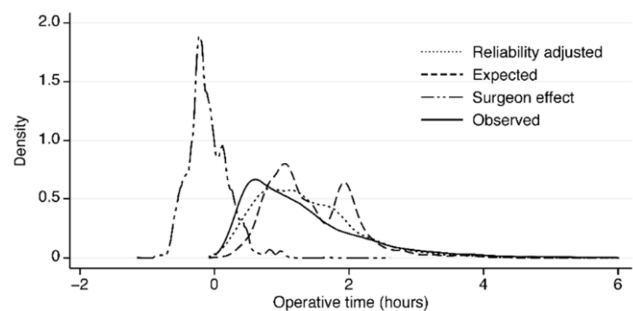
### 2.8. Identification of a Learning/Improvement Effect

Evidence of learning requires measurable improvement over time. A second larger dataset of 171,116 cases was used to establish the presence of a learning/improvement effect.

Each postoperative occurrence was tested as an improvement measure over the 12 years for which there is data, using the three-level, mixed-effects model with a variable for year. The measure is the coefficient or odds ratio (95% confidence interval) for the independent variable year, dependent upon regression type.

All analyses were performed using 64-bit STATA/MP 16.1 for Windows (College Station, Texas, U.S.A.). The modeling methods proposed by Rabe-Hesketh and Skrondal were followed [13].

## 3. Results



**Figure 1.** Shows the density distributions of the observed and risk model reliability-adjusted outputs, i.e., expected and surgeon effects for OT. Like a frequency distribution, the density peaks when the number of observations peak. The reliability-adjusted OT is the sum of the fixed and random effects (hospital and surgeon). The expected OT represents the fixed effects only. The surgeon effects show the surgeon effect (the surgeon random intercept) on OT in this data set and ranged from -1.2 hours to 9.4 hours.

The surgeon effects represent the surgeon's impact on the true score; in this example, OT. The hospital and surgeon signals, which are expressed as the mean and standard deviation, and the range of the surgeon effects are shown in Table 1. The hospital signal is zero for the three candidate measures of PE, CVA, and UI. The surgeon signal is zero for the six candidates: CPR, DVT, mortality, MI, PE, and SHOCK. The range of the surgeon effects (random intercepts) is zero for all candidate measures in which the signal is zero and PRI. The surgeon effects are expressed in probability units for all candidate measures, except for OT, which is reported in hours, and LOS, which is reported in days. The surgeon signal did not show majority control; thus, the hospital signal was larger than the surgeon signal for eleven postoperative occurrences: mortality, CPR, DSSI, DVT, MI, OSSI, PNA, PRI, sepsis, SHOCK, and SSSI.

**Table 1.** Hospital and Surgeon signals and surgeon effects results.

Candidate Measure	Between variance (signals)		Case-Weighted Surgeon Random Intercepts				Measure
	Hospital	Surgeon	Mean	Std. Dev.	Min	Max	Units
Acute Renal Failure	0.0247	0.1289	-0.0191	0.1142	-0.3258	0.4924	Probability
Bleeding/Transfusions	0.0623	0.2569	-0.0123	0.2885	-0.7866	1.0358	Probability
Cardiac Arrest Requiring CPR	0.0324	0.0000	0.0000	0.0000	0.0000	0.0000	Probability
Deep Incisional SSI	0.7850	0.3056	-0.0120	0.1421	-0.4334	1.0852	Probability
Deep Venous Thrombosis	0.1563	0.0000	0.0000	0.0000	0.0000	0.0000	Probability
Length of Stay	0.0141	0.0948	0.0141	0.1059	-0.1830	0.1574	Days

Candidate Measure	Between variance (signals)		Case-Weighted Surgeon Random Intercepts				Measure
	Hospital	Surgeon	Mean	Std. Dev.	Min	Max	Units
Mortality	0.0680	0.0000	0.0000	0.0000	0.0000	0.0000	Probability
Myocardial Infarction	0.1395	0.0000	0.0000	0.0000	0.0000	0.0000	Probability
On Ventilator > 48 Hours	0.0221	0.1290	-0.0036	0.1144	-0.3314	0.5132	Probability
Operative Time	0.0058	0.1848	-0.0660	0.3605	-1.1995	9.3912	Hours
Organ/Space SSI	0.3058	0.1280	-0.0018	0.1147	-0.2254	0.5103	Probability
Patients With Morbidity	0.0674	0.1024	-0.0024	0.1836	-0.5637	0.5563	Probability
Pneumonia	0.3392	0.0026	-0.0002	0.0109	-0.0375	0.0571	Probability
Progressive Renal Insufficiency	0.2339	0.0246	0.0000	0.0000	0.0000	0.0000	Probability
Pulmonary Embolism	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	Probability
Return to Operating Room	0.0150	0.1483	0.0043	0.1864	-0.4451	0.6748	Probability
Sepsis	0.1608	0.1379	-0.0088	0.1058	-0.2996	0.4057	Probability
Septic Shock	0.3111	0.0000	0.0000	0.0000	0.0000	0.0000	Probability
Stroke/CVA	0.0000	0.0438	-0.0012	0.0159	-0.0700	0.0721	Probability
Superficial Incisional SSI	0.3891	0.1024	0.0028	0.1098	-0.3177	0.5002	Probability
Unplanned 30-Day Readmissions	0.0268	0.0432	0.0053	0.0812	-0.1574	0.3386	Probability
Unplanned Intubation	0.0000	0.0008	-0.0019	0.0201	-0.0725	0.0801	Probability
Urinary Tract Infection	0.0940	0.2147	-0.0018	0.1608	-0.3267	0.6483	Probability
Wound Disruption	0.0000	0.7593	-0.0764	0.3470	-0.8442	1.3310	Probability

Surgeon effects range from zero for seven candidate measures to -0.8 to 1.33 probability units for wound disruption. Length of stay, surgeon effects, range from -0.18 to 0.16 days. Operative time surgeon effects range from -1.2 hours to 9.4 hours.

**Table 2.** Likelihood ratio test results of the hypothesis that the random intercepts do not exist.

Candidate Measure	Comparison to standard regression*		Surgeon intercept exists**	
	Test	P > $\chi^2$	Test	P > $\chi^2$
Length of Stay	1474.42	0.0000	1038.41	0.0000
Operative Time	2812.85	0.0000	2350.60	0.0000
Mortality	6.89	0.0043	N/A <sup>†</sup>	
Acute Renal Failure	0.31	0.8558	0.20	0.3274
Bleeding/Transfusions	89.13	0.0000	52.00	0.0000
Cardiac Arrest Requiring CPR	0.15	0.3488	N/A <sup>†</sup>	
Deep Incisional SSI	40.12	0.0000	1.65	0.0993
Deep Venous Thrombosis	8.65	0.0016	N/A <sup>†</sup>	
Myocardial Infarction	3.44	0.0319	N/A <sup>†</sup>	
On Ventilator > 48 Hours	3.05	0.2176	2.04	0.0768
Organ/Space SSI	19.24	0.0001	1.73	0.0940
Pneumonia	59.90	0.0000	0.01	0.4711
Progressive Renal Insufficiency	6.08	0.0477	0.02	0.4467
Pulmonary Embolism	0.00	N/A <sup>‡</sup>	N/A <sup>†</sup>	
Sepsis	21.84	0.0000	1.58	0.1044
Septic Shock	15.65	0.0000	N/A <sup>†</sup>	
CVA/Stroke	0.04	0.4230	0.04	0.4230
Superficial Incisional SSI	98.51	0.0000	2.31	0.0618
Unplanned Intubation	0.00	0.4957	0.00	0.4957
Urinary Tract Infection	7.72	0.0210	3.31	0.0230
Wound Disruption	7.22	0.0036	N/A <sup>†</sup>	
Patients With Morbidity	103.43	0.0000	40.97	0.0000
Unplanned 30-Day Readmissions	21.45	0.0000	5.21	0.0113
Return to Operating Room	32.70	0.0000	21.71	0.0000

\* The likelihood ratio test comparing the three-level model that includes random effects for both hospital and surgeon to a standard regression.

\*\* The likelihood ratio test comparing the three-level model that includes both the hospital and surgeon random effects to the same model with the surgeon variance set to zero.

<sup>†</sup> This comparison could not be tested for this postoperative occurrence model because the between-surgeon variance in both models is zero.

<sup>‡</sup> This comparison could not be tested because both the between-hospital and between-surgeon variances were zero.

Table 2 shows the likelihood ratio test results in which the three-level risk models were not helpful over a standard regression analysis for five postoperative occurrences: ARF, CPR, ONVENT, CVA, and UI. The hypothesis that a random

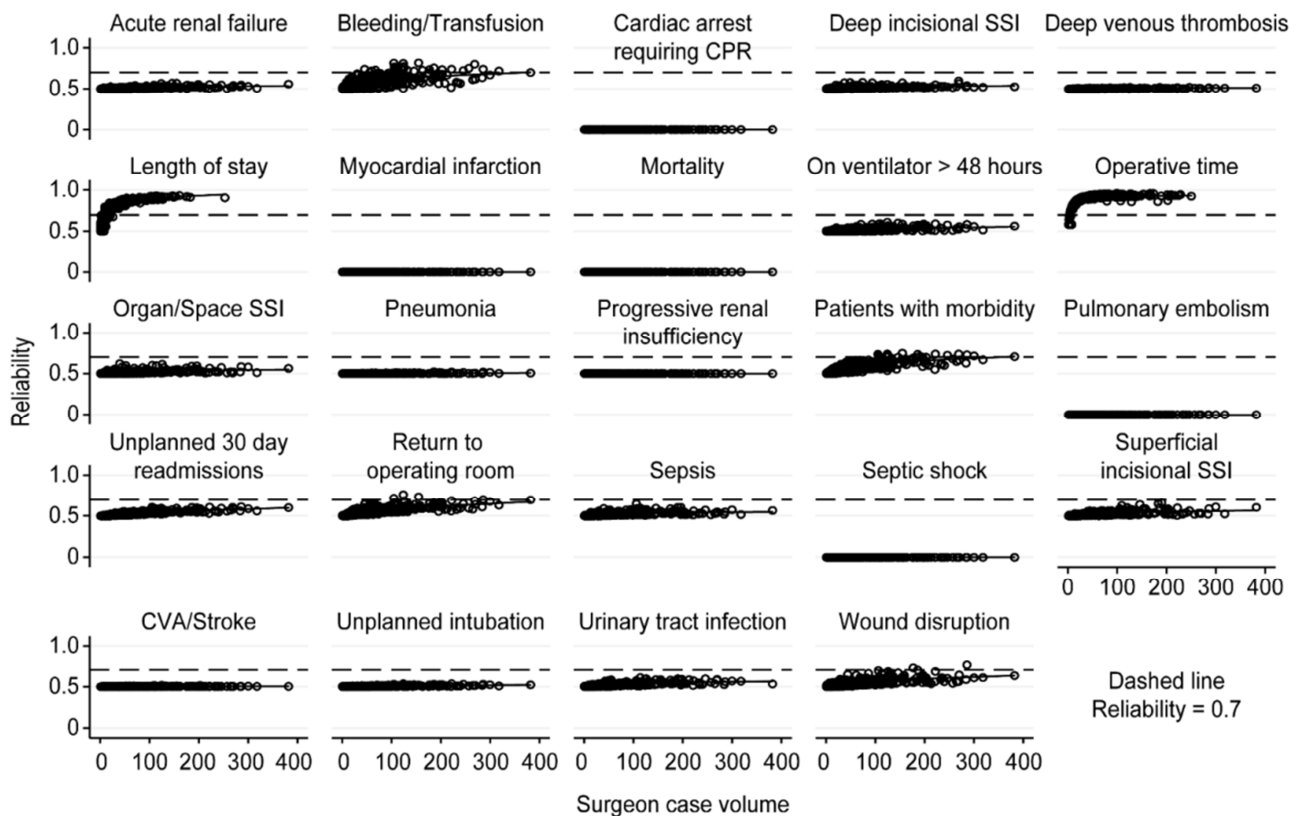
surgeon intercept does not exist was true for ten risk models: ARF, DSSI, ONVENT, OSSI, PNA, PRI, sepsis, CVA, SSSI, and UI.

The associations between the candidate measures and other

quality measures (all candidates) are reported in the appendix. All candidate measures were associated with the other candidate measures, ranging from a low of 4 measures for CPR to a high of 20 measures for BT. The presence of a learning/improvement effect was confirmed in ten of the candidate measures: in operative time the annual improvement (coefficient) was  $-0.007$ ;  $P < 0.0005$ ; ARF the odds ratio (OR) was  $0.944$  ( $0.90 - 0.99$ ); DSSI, OR =  $0.90$  ( $0.86 - 0.94$ ); ONVENT, OR =  $0.90$  ( $0.88 - 0.93$ ); SSSI, OR =  $0.93$  ( $0.01 - 0.96$ ); SEPSIS, OR =  $0.92$  ( $0.889 - 0.94$ ); SEPTIC SHOCK, OR =  $0.91$  ( $0.87 - 0.94$ ); UTI, OR =  $0.90$  ( $0.88 - 0.93$ ); WD, OR =  $0.91$  ( $0.87 - 0.95$ ) and ROR, OR =  $0.94$  ( $0.93 - 0.96$ ). Three measures increased/worsened over time BT, OR =  $1.02$  ( $1.00 - 1.04$ ); MI, OR =  $1.05$  ( $1.00 - 1.09$ ); READS, OR =  $1.27$  ( $1.26 - 1.30$ ).

The reliability of each surgeon candidate measure versus surgeon case volume is shown in Figure 2. No surgeons met the reliability criterion of  $0.70$  for eighteen of the twenty-four measures: mortality, ARF, CPR, DSSI, DVT, MI, ONVENT, OSSI, PNA, PRI, PE, sepsis, SHOCK, SSSI, CVA, UI and UTI. Twenty-eight (4.3%) surgeons met the reliability criterion for BT, 375 (58.2%) surgeons met the reliability criterion for LOS, 527 (81.8%) surgeons met the reliability criterion for OT, 15 (2.3%) surgeons met the reliability criterion for PTSWMB, 4 surgeons met the reliability criterion for ROR and 2 surgeons met the reliability criterion for WD. The classification of the surgeons into the high-risk and low-risk categories is shown in Table 3. A high risk of a postoperative occurrence could be detected in 8 of the 24

candidate measures: LOS, OT, BT, DSSI, WD, PTSWMB, READ, and ROR. A low risk of a postoperative occurrence could be detected in 4 of the 24 candidate measures: LOS, OT, BT, and PTSWMB. In total, 112 (17.4%) surgeons were identified with a low risk of a postoperative occurrence, and 139 (21.6%) surgeons were identified with a high risk. Four of the 24 candidate measures identified surgeons with both high and low risk. Neither high- nor low-risk surgeons could be identified in 16 of the candidate measures: mortality, ARF, CPR, DVT, MI, ONVENT, OSSI, PNA, PRI, PE, sepsis, SHOCK, CVA, SSSI, UI, and UTI. The surgeon risk classification in BT and PTSWMB was low; in the BT measure, two low-risk surgeons and five high-risk surgeons (1.4% of the surgeons) were identified, and in the PTSWMB measure, one low-risk and five high-risk surgeons (0.94% of the surgeons) were identified. The OT and LOS measures classification was 26.3% and 12.3% of the assessed surgeons, respectively. The sensitivity analysis of surgeon reliability in the PTSWMB measure shows that only 15 surgeons met the  $0.7$  reliability threshold in the three-level model. In comparison, 52 surgeons met the threshold in the two-level model. The model result differences contributing to the better reliability performance included a 42% larger between-surgeon variance ( $0.1447$  versus  $0.1018$ ), and the surgeon random intercept variance was greater by 22% ( $0.081$  versus  $0.066$ ). Comparing the two models, the likelihood ratio test was  $18.32$ ,  $P < 0.0001$ , indicating that the three-level model with both hospital and surgeon random intercepts is better than the model with only the surgeon random intercepts.



**Figure 2.** Reliability of surgeon candidate measures (random intercepts of postoperative occurrences).

**Table 3.** Surgeon classification results: Surgeons with random intercepts significantly lower or greater than the case-averaged intercept.

Candidate Measure	RISK		Total	Total	% Surgeons <	% Surgeons >
	LOW	HIGH	Cases	Surgeons	Case Averaged Intercept	
Length of Stay	30	42	15366	586	5.12	7.17
Operative Time	79	77	19412	593	13.32	12.98
Bleeding/Transfusions	2	7	29267	644	0.31	1.09
Deep Incisional SSI	0	1	29267	644	0.00	0.16
Wound Disruption	0	2	29267	644	0.00	0.31
Patients With Morbidity	1	5	29267	644	0.16	0.78
Unplanned 30-Day Readmissions	0	1	29267	644	0.00	0.16
Return to Operating Room	0	4	29267	644	0.00	0.62
Grand Total	112	139			18.91	23.26

The following sixteen measures did not detect either the high- or low-risk surgeons: mortality, ARF, CPR, DSSI, DVT, MI, ONVENT, OSSI, PNA, PRI, PE, sepsis, SHOCK, CVA/stroke, SSSI, UI and UTI.

## 4. Discussion

The establishment of a high-quality registry of clinical information for surgical cases and outcomes facilitates quality improvement efforts [14]. In addition to the LOS and OT, surgical case mortality and morbidity have been proposed as measures for the assessment of the quality of surgical intervention. The LOS has been promoted as a quality measure by the Committee on Trauma of the ACS and has been positively impacted by Enhanced Recovery After Surgery protocols [15, 16]. The duration of surgery has previously been used as a quality measure in the United Kingdom [17-19].

In this three-level random intercept model of patient risk, the sum of the fixed ( $\beta$ ) and random effects ( $\beta x_i + u_j + u_k$ ) is the true risk score for patient  $i$ , with hospital  $k$ , surgeon  $j$ , and risk, demographic and case-mix factors,  $x$ . The random intercept model shifts the overall regression line according to each surgeon and hospital, but the slope,  $\beta$ , remains constant. The random effect (random intercept) of surgeon  $j$  represents the individual differences compared to other surgeons due to personal characteristics that are not included as variables in the model. Since neither  $\beta$  nor  $x$  varies by surgeon and  $u_k$  varies by hospital, exploring potential measures of surgeon performance required comparing  $u_j$ , which is the surgeon random intercept among surgeons  $u_1$  through  $u_{644}$  in the current study for most assessed measures (LOS and OT provided results for 586 and 593 surgeons, respectively). The fixed effects, or the "slope" of the model, are important in developing surgeon random intercepts that are properly adjusted for covariates' presence. The fixed effects do not add to the surgeon's performance assessment once the intercepts have been estimated and do not impact the surgeon ranking. The surgeon random intercept estimation is adjusted by the inclusion of the fixed effects representing the patient demographics, preoperative risk, case-mix factors, and hospital random effects. A larger surgeon random intercept indicates that for the same fixed-effects result, a patient's risk for an increased LOS, for example, is greater.

The patient-level "true score" measure includes the surgeon and hospital random effects plus the fixed effects in the numerator of the incident rate ratio, creating a measure of surgeon performance that is confounded by the hospital effects. Both the hospital and surgeon effects range from negative to positive for each candidate measure (Table 1);

controlling for the between-hospital variance in the estimation of the surgeon performance by using a three-level model reduces the error associated with an ambiguous performance measure. Surgeon performance can be estimated using an incident rate ratio, where the numerator is the surgeon random effects plus fixed effects and the denominator is the fixed effects. However, because surgeon performance is measured using a random intercept, a comparison to the population-averaged intercept is intuitively more appealing and eliminates the counterintuitive comparison to the fixed effects, which include the patient demographics, patient preoperative risks and procedure risks, all of which may influence but are not measures of surgeon performance.

Assessing the performance of the three-level model by testing for random intercepts at both levels 2 (surgeon) and 3 (hospital) is helpful and reduces the potential error of using model results where no surgeon random intercept is present as a measure of surgeon performance. In this experience using this dataset, only seven of the 24 candidate measures have surgeon random intercepts. As measured in this study, majority surgeon control prevents holding the surgeon accountable for an outcome that has historically been controlled by the hospital. Surgeons who work in a three-level system at more than one level (as many do) may have an opportunity to influence policy at the hospital level and, consequently, can play a role in improving a target measure that is traditionally not influenced by surgeons. The candidate measures with no surgeon random intercept, i.e., no surgeon effect on their outcome, could require a non-traditional approach to establish a surgeon effect. In contrast, surgeons have a large effect on OT and can use technologies, such as robotic or other forms of minimally invasive surgery, to mitigate the impact of long procedure durations. Finally, surgeons can also influence the OT through additional learning and experience, as shown in this study.

The prevalence of postoperative morbidity and the distribution of cases among surgeons do not favor the use of morbidity as a measure of surgeon performance. Thirty-two percent of surgeons have fewer than 10 cases; eight percent of surgeons have only one case. Only 10 percent of surgeons have patients with DSSI, while 30 percent of surgeons have no patients with postoperative morbidity. WD has the largest

surgeon signal, but in the 104 events, only two surgeons met the reliability threshold of 0.7. In total, 3,288 of 29,267 cases had a(ny) morbidity, and only 15 surgeons met the reliability threshold for this measure. The second most prevalent morbidity is transfusion, with 1,483 cases, and only 28 surgeons met the reliability threshold. Shih et al. concluded that when assessing the colectomy complication rates, statistical noise, as evaluated by low reliability, is a significant determinant of most surgeons' surgeon-specific complication rate due to the low case volume. Hall et al. reported that 61.9 percent of surgeons achieved a reliability of 0.7 for their morbidity measure. However, Hall et al. did not control for the between-hospital variance of morbidity because they used only a two-level model, including the surgeon and patient. The hospital level's exclusion from their model of postoperative occurrences created the potential for a confounded between-surgeon variance and an inflated estimate of reliability. Postoperative process measures, such as RORs and READs, also suffer from a low prevalence, uneven case distribution, and low reliability. No surgeons met the minimum reliability threshold of 0.7 for unplanned READs within 30 days of discharge (1,605 events). Only 4 surgeons met the reliability threshold for ROR (1,101 events). In contrast, 104,799 inpatient days and 47,258 operative hours were reported. Three hundred seventy-five (58.4%) surgeons met the reliability threshold for LOS, and 527 (81.8%) surgeons met the reliability threshold for OT. Only 34 of 644 (5.3%) surgeons had no inpatient days, and all surgeons have OT.

The generalizable results of this study include several important points: OT is an excellent surgeon performance measure, while most postoperative outcome measures are limited by a low prevalence, no or low surgeon control or an inability to classify risk. The LOS is a good surgeon performance measure, while BT and PTSWMB may be used selectively but lack the characteristics to be widely applicable. Careful consideration of the surgeon signal's presence and magnitude provides insight into the possible mechanisms by which reductions in postoperative occurrences can be achieved and whether the primary vector occurs at the hospital or surgeon level. The intraclass correlation could be used to determine the relative level of surgeon control in linear, logistic, and probit models. Due to the multilevel mixed-effects negative binomial model for LOS, and the desire to compare the control levels across model types, evaluation of the surgeon signal was used. The study results that are unlikely to be generalizable include the surgeon and hospital signals because care approaches may vary geographically and over time. However, this lack of generalizability also presents an opportunity for further studies to explore how the most significant surgeon effect can be achieved by examining the varied approaches to care for each postoperative occurrence.

## 5. Conclusions

Comparing outcomes across surgeons differs from measuring surgeon performance. Comparing outcomes requires a risk adjustment and the use of the "true score"

approach. Still, it is limited by the constraints of case volume and a confounding factor, i.e., the hospital, if used to judge surgeons' performance. Assessing surgeon performance requires a measure of the surgeon's effects on the consequences (postoperative occurrences) of surgical procedures, i.e., the surgeon-specific random intercept, which is a product of a multilevel risk adjustment model. Postoperative morbidities and mortality lack the characteristics necessary to be used as measures of surgeon performance. The combination of low prevalence rates, low case numbers, low reliability, and limited ability to classify surgeons by risk generally precludes their use. The postoperative measures of process, ROR, and READs are also affected by low prevalence rates, low case numbers, and low reliability at the surgeon level. However, the process measures LOS and OT both have high event rates and high reliability. Controlling for the between-hospital variance of the postoperative occurrence in a three-level model reduces the probability of the hospital's influence on the candidate measure of surgeon performance. There is no control for between-hospital variance in a two-level model, and the surgeon reliability may be artifactually higher. Improvement or learning effects enhance the appeal of measures for evaluating surgeon performance.

## References

- [1] Yi, S. G., N. P. Wray, S. L. Jones, B. L. Bass, J. Nishioka, S. Brann, and C. M. Ashton. 2013. "Surgeon-Specific Performance Reports in General Surgery: An Observational Study of Initial Implementation and Adoption." *Journal of the American College of Surgeons* 217 (4): 636-647.E631.
- [2] Allen, M., and O. Pierce. 2015. "Making the Cut: Why Choosing the Right Surgeon Matters Even More than You Know." ProPublica, July 13, 2015.
- [3] Birkmeyer, N. J., and J. D. Birkmeyer. 2006. "Strategies for Improving Surgical Quality--Should Payers Reward Excellence or Effort?" *New England Journal of Medicine* 354 (8): 864-870.
- [4] Cooper, W. O., O. Guillaumondegui, O. J. Hines, C. S. Hultman, R. R. Kelz, P. Shen, D. A. Spain, J. F. Sweeney, I. N. Moore, J. Hopkins, I. R. Horowitz, R. M. Howerton, J. W. Meredith, N. O. Spell, P. Sullivan, H. J. Domenico, J. W. Pichert, T. F. Catron, L. E. Webb, R. R. Dmochowski, J. Karrass, and G. B. Hickson. 2017. "Use of Unsolicited Patient Observations to Identify Surgeons with Increased Risk for Postoperative Complications." *JAMA Surgery* 152 (6): 522-529.
- [5] Stephens, M. R., W. G. Lewis, A. E. Brewster, I. Lord, G. R. Blackshaw, I. Hodzovic, G. V. Thomas, S. A. Roberts, T. D. Crosby, C. Gent, M. C. Allison, and K. Shute. 2006. "Multidisciplinary Team Management is Associated with Improved Outcomes after Surgery for Esophageal Cancer." *Diseases of the Esophagus* 19 (3): 164-171.
- [6] Nurok, M., N. Sadovnikoff, and B. Gewertz. 2016. "Contemporary Multidisciplinary Care--Who Is the Captain of the Ship, and Does It Matter?" *JAMA Surgery* 151 (4): 309-310.

- [7] Statement on Physician-Led Team-Based Surgical Care. 2016. "American College of Surgeons." [accessed on August 11, 2017]. Available at: <https://www.facs.org/about-acsc/statements/86-team-based>.
- [8] Hall, B. L., K. M. Huffman, B. H. Hamilton, J. L. Paruch, L. Zhou, K. E. Richards, M. E. Cohen, and C. Y. Ko. 2015. "Profiling Individual Surgeon Performance Using Information from a High-Quality Clinical Registry: Opportunities and Limitations." *Journal of the American College of Surgeons* 221 (5): 901-913.
- [9] Shih, T., A. I. Cole, P. M. Al-Attar, A. Chakrabarti, H. A. Fardous, P. F. Helvie, M. T. Kemp, C. Lee, E. Shtull-Leber, D. A. Campbell, M. J. Englesbe. 2015. "Reliability of Surgeon-Specific Reporting of Complications after Colectomy." *Annals of surgery* 261 (5): 920-925.
- [10] Iezzoni, L. I. 2003. *Risk Adjustment for Measuring Health Care Outcomes*, Third Edition. Chicago, Illinois: Health Administration Press.
- [11] Adams, J. L. 2009. *The Reliability of Provider Profiling: A Tutorial*. Santa Monica: RAND Corporation.
- [12] Adams, J. L., A. Mehrotra, J. W. Thomas, and E. A. McGlynn. 2010. "Physician Cost Profiling — Reliability and Risk of Misclassification." *New England Journal of Medicine* 362 (11): 1014-1021.
- [13] Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*, Volumes I and II, Third Edition. College Station, TX: Stata Press. Stata Corp LP.
- [14] Guillaumondegui, O. D., O. L. Gunter, L. Hines, B. J. Martin, W. Gibson, P. C. Clarke, W. T. Cecil, and J. B. Cofer. 2012. "Using the National Surgical Quality Improvement Program and the Tennessee Surgical Quality Collaborative to Improve Surgical Outcomes." *Journal of the American College of Surgeons* 214 (4): 709-714.
- [15] Rotondo, M. F., C. Cribari, and R. S. Smith. 2014. *Resources for Optimal Care of the Injured Patient*. Chicago, Illinois.: Committee on Trauma, American College of Surgeons.
- [16] Stone, A. B., M. C. Grant, C. Pio Roda, D. Hobson, T. Pawlik, C. L. Wu, and E. C. Wick. 2016. "Implementation Costs of an Enhanced Recovery after Surgery Program in the United States: A Financial Model and Sensitivity Analysis Based on Experiences at a Quaternary Academic Medical Center." *Journal of the American College of Surgeons* 222 (3): 219-225.
- [17] Hopper, A. N., M. H. Jamison, and W. G. Lewis. 2007. "Learning Curves in Surgical Practice." *Postgraduate Medical Journal* 83 (986): 777-779.
- [18] Jaffer, A., B. Bednarz, B. Challacombe, and S. Sriprasad. 2009. "The assessment of surgical competency in the UK." *International Journal of Surgery* 7 (1): 12-15.
- [19] Khan, N., H. Abboudi, M. S. Khan, P. Dasgupta, and K. Ahmed. 2014. "Measuring the surgical 'learning curve': methods, variables and competency." *BJU International* 113 (3): 504-508.