A New Composite Measure for Assessing Surgical Performance

Rahul S Iyengar, MD; James M Mossner, BS; Shaina Sekhri, MD;
Andrew Mullard, MS; Greta Krapohl, PhD, RN;
Darrell A Campbell Jr, MD; Michael J Englesbe, MD

ABSTRACT

Background

Surgeon scorecards can improve transparency and empower patients to become informed consumers, but previous efforts have fallen short. In response, we have developed the Colectomy Scorecard, a composite measure that utilizes high-quality clinical registry data to assess surgeon performance on colectomies across five domains: survival, morbidity avoidance, anastomotic leak avoidance, SSI bundle compliance, and utilization. We seek to compare our measure with ProPublica's Surgeon Scorecard for reliability and agreement in predicting future surgeon performance.

Methods

This study used surgeon-specific clinical registry data from the Michigan Surgical Quality Collaborative (MSQC) generated from elective colectomy cases over a 30-month time period ($n = 7032$). Our model utilized 34 weighted items across 5 domains to develop a composite score for each surgeon,

and ratings were assigned to low- and high-performing outliers. Statistical reliability and test-retest agreement were assessed using actual and simulated data sets. The analytic approach from ProPublica's Surgeon Scorecard was also applied utilizing MSQC data, and test-retest agreement was compared between the two models.

Results

Our scorecard predicts future performance for low- and high-performing surgeons with 89.0% accuracy ($p < 0.01$, 95% CI 88.3–93.2), suggesting that the addition of clinical domains can strengthen the measurement of the surgical episode. Furthermore, our model achieves a reliability of 0.80 (95% CI 74.8–84.9) and 0.90 (95% CI 87.9–93.3) for surgeon colectomy case volumes of 62 and 170, respectively.

Conclusions

We have developed a preliminary composite Colectomy Scorecard

Author E-mail List:

iyengarr@med.umich.edu;
jamemoss@med.umich.edu;
ssekhri@med.umich.edu;
andrew.mullard@gmail.com;
krapohl@med.umich.edu;
darrellc@med.umich.edu;
englesbe@med.umich.edu

Corresponding Author:

Rahul Iyengar, BSE,
iyengarr@med.umich.edu,
248-909-6433

Conflicts of Interest:

The authors have no conflicts of interest to disclose at this time.

that reliably predicts future surgeon performance. By providing domain-specific feedback to surgeons, the Colectomy Scorecard outlines a

framework that can be adapted to other surgical specialties to help hospitals improve transparency and drive outcomes improvement.

Introduction

Within the medical community is a strong desire to increase transparency between healthcare providers and patients. Achieving transparency in healthcare requires making available information about the cost and quality of healthcare services so that patients can become informed consumers.¹ Transparency increases trust and improves dynamics between patients and physicians when complete, objective, and high-quality data regarding patient care is made available.^{2,3} To achieve this ambitious goal, the measures used for public reporting must be accepted by all stakeholders and must reliably predict the performance of the physicians being assessed.

One strategy for improving transparency is to utilize scorecards to evaluate surgeon performance.^{1,4,5} One of the most well-known and controversial surgeon scorecards released to date is ProPublica's Surgeon Scorecard.⁶ ProPublica is an independent, non-profit newsroom that released their scorecard in July 2015. Although this scorecard was a novel approach to increasing transparency between patients and their surgeons by directly releasing scores to the public, it was met with much criticism from the medical community.⁷⁻¹¹ ProPublica's scorecard was criticized for its use of administrative Medicare claims billing data filled with inconsistencies and a composite score composed of unreliable metrics.^{8,10,11}

Using robust clinical registry data from the Michigan Surgical Quality Collaborative (MSQC) and healthcare claims data from Blue Cross and Blue Shield of Michigan (BCBSM), we have developed a composite Colectomy Scorecard to include components that are important to patients, surgeons, and payers.¹² The blend of administrative and clinical metrics that we utilize to grade surgeons performing colectomies are survival, morbidity

avoidance, anastomotic leak avoidance, surgical site infection (SSI) bundle compliance, and utilization (costs of care). These metrics are combined to generate a composite score for each MSQC surgeon, allowing for greater reliability in comparison to individual measures.^{9,13} We hypothesize that our multi-domain scorecard approach can provide strong statistical reliability and accurately predict future surgeon performance.

Methods

Data Source

This study utilizes clinical registry data from the MSQC. Funded by the Blue Cross and Blue Shield of Michigan, the MSQC is a coalition of 72 academic and community hospitals voluntarily engaged in data collection, feedback, and collaborative quality improvement in the state of Michigan.^{12,14} At each participating hospital, specially trained, dedicated nurse abstractors collect patient characteristics, intraoperative processes of care, and 30-day postoperative outcomes from general and vascular surgery and hysterectomy cases in accordance with established policies and procedures. To reduce sampling error, a standardized data collection methodology is employed that uses select cases on an 8-day cycle (thus alternating weekdays sampled for each cycle). The clearly defined and standardized data collection methodology is routinely validated through scheduled site visits, conference calls, and internal audits.

Study Data

Cases included for this scorecard were colectomy procedures as identified by multiple principal Current Procedural Terminology (CPT) codes charted in electronic health records; these codes are listed in Table 1—*online only*. We selected for operation dates between July 2012 and April 2015. Only patients aged 18 years or older were included

in the study. In total, 7032 cases performed by 469 surgeons across 64 hospitals were included in this study. We did not have formal exclusion criteria beyond cases that fell outside of these aforementioned parameters. Given the source of our data—clinical registry maintained by trained nurse abstractors—missing data for surgical cases was not an issue, as patients with missing information were not included in the accessible registry pool. Cases that were revisions or following readmissions are captured by the items in the utilization domain present in Table 2.

Scorecard Domains and Items

The Colectomy Scorecard consists of a multidimensional measure of overall surgical quality. The scorecard consists of 34 items under 5 domains: (1) survival, (2) morbidity avoidance, (3) anastomotic leak avoidance, (4) SSI bundle compliance, and (5) utilization (Table 2). These domains and the items that comprise them were selected following a survey of surgeons from the MSQC, based on their clinical expertise.

We subdivided our 5 domains into core domains (applicable to any surgical field) and procedure-specific domains (applicable specifically to colectomies). Mortality and morbidity are common outcome measurements utilized extensively to assess quality across surgical specialties; they are encompassed by the survival and morbidity avoidance core domains. Anastomotic leak avoidance is represented as a procedure-specific domain, given the particular prevalence and impact of anastomotic leak on colectomy patients.¹⁵ SSI bundle compliance refers to compliance with

6 process measures recommended by the MSQC that are associated with improved quality and value of surgical care in colectomy patients, thus representing a procedure-specific domain.^{16,17} Finally, the utilization core domain reflects the consumption of additional postoperative healthcare resources associated with sub-optimal outcomes and increased costs.

Surgeon Scorecard Construction

To generate a composite scorecard, case-mix adjustment models were first fitted for each of the 34 contributing items across the 5 domains. For each of these contributing items, we created a case-mix adjustment model for multiple patient-level variables to support a balanced analysis that does not unfairly punish surgeons performing complex cases. The domains, contributing items, and patient-level variables used for these models are outlined in Table 2. Each case-mix adjustment model generated an output variable with a Pearson residual for every item. By adjusting for the patient-level factors, the Pearson residual is thus representative of individual surgeon variation and performance.^{18,19}

Next, the Pearson residuals were rescaled to a 0 to 1 scale in order to equalize the contribution of each item. Weights were then applied to these rescaled Pearson residuals, and composite measure scores were generated by summing the weighted and rescaled Pearson residuals across domains. These weightings created a total scale of 0 to 100 and are outlined in Table 2. We balanced a combination of patient-centered clinical outcomes, process measure compliance,

TABLE 1. ONLINE ONLY—CPT Codes for Case Inclusion Criteria

Note: The following patient variables were incorporated when analyzing each of the above 34 items: age, gender, BMI, ASA physical status classification, independent functional status, comorbidity count, diabetes, smoking history, alcohol history, dyspnea, ventilator use, COPD, pneumonia, sleep apnea, ascites, cirrhosis, CHF, arrhythmias, coronary artery disease, hypertension, beta blocker use, statin use, peripheral vascular disease, DVT, family history of DVT, dialysis, sepsis, urinary tract infections, cancer, presence of open wound, chronic steroid use, body weight loss, bleeding disorder, history transfusion, creatinine, albumin, bilirubin, blood glucose, hemoglobin, WBC, HCT, platelet count, lactate, INR, acuity, wound classification, ostomy, and operative duration.

TABLE 2. Domains, Items, and Weights for the Colectomy Scorecard

Domain Type	Domain Name	Items	Item Count	Item Weight	Weighted Total
Core	[A] Survival	• Mortality	1	10	10
Core	[B] Morbidity Avoidance	• Acute Renal Insufficiency/Failure • Pneumonia • Sepsis • Severe Sepsis • Surgical Site Infection—Superficial • Surgical Site Infection—Organ Space • Surgical Site Infection—Deep • UTI-SUTI • UTI-CAUTI • CLABSI • Clostridium Difficile • DVT req. Therapy • Pulmonary Embolism • Stroke/CVA • Cardiac Arrest requiring CPR (2 items) • Myocardial Infarction (2 items) • Cardiac Arrhythmias • Intubation (2 items) • Transfusion 72hrs	22	1	22
Procedure Group Specific	[C] Anastomotic Leak Avoidance	• Anastomotic Leak	1	10	10
Procedure Group Specific	[D] SSI Bundle Compliance	• Appropriate IV Prophylactic Antibiotics • Postoperative Normothermia (temp>98.6°F) • Oral Antibiotics with Mechanical Bowel Prep • Postoperative Day 1 Glucose ≤140mg/dL • Minimally Invasive Surgery • Short Operative Duration (<100min)	6	5	30
Core	[E] Utilization	• Extended Hospital LOS • ED Visit • Readmission • Reoperation	4	7	28
Mixed	[A,B,C,D,E] Colectomy Quality Measure	-	34	-	100

and value measures in conjunction with recommendations from MSQC surgeons to determine these weights. For example, rare, life-threatening adverse events have a public health and financial impact that is disproportionately large. However, weighting rare events heavily tends to strongly decrease repeatability and require excessively large sample sizes.²⁰ Thus, it was necessary to balance these considerations to produce a quality measure that is of practical use. Finally, the distribution of weighted scores was normalized using a rank-based inverse normal transformation to improve model fit.²¹

Characterizing the Variable Type of the Composite Measure

While the composite score is on an ordinal scale with 34 levels, combining these levels with the usage of Pearson residuals and normal transformation produces data nearly equivalent to a continuous scale. The distribution of total composite scores closely approximates a normal distribution (Figure 1—*online only*). These 2 observations support analysis of the composite score as a continuous response variable.^{22–24} The use of a composite score, rather than 5 separately reported domains, allows for improved statistical power and reliability.^{9,13,25}

Provider Profiling and Star Rating

We predicted surgeon performance by fitting the scores to a Bayesian linear random-effects model. This model assumes that surgical errors are independent and identically distributed within a normal distribution.^{18,19,26} With this model, surgeon outliers that deviated from expected performance were identified by 95% confidence intervals that did not overlap the overall mean. Surgeons were assigned ratings of 1 star for low performing outliers, 2 stars to middle tier non-outliers, and 3 stars for high-performing outliers.

Reliability and Agreement

The discrimination and reproducibility of measurements for the surgeon scorecard were evaluated using reliability and agreement. Reliability was used to show how well scores discriminate between surgeons, despite measurement error, whereas agreement quantified how often surgeon star ratings were identical on repeated measurement.^{27,28}

Reliability (λ) coefficients range from 0 to 1 where 0 represents no reliability and 1 represents perfect reliability. We characterized reliability of $\lambda \geq 0.8$ as good and $\lambda \geq 0.9$ as excellent. Reliability of the surgeon was estimated as a random effect. With this approach, reliability is a function of the intra-class correlation for surgeons and the number of cases sampled per surgeon.^{19,25,26,29,30} This allows us to determine the case volumes necessary to achieve different levels of reliability for our presented scorecard methodology.

Agreement was measured as the proportion of surgeons with identical outlier star ratings on repeated measurement. This represents calculating the raw overall agreement of star ratings from 2 samples, with multiple star rating measures. In essence, this approach assesses the ability of the scorecard to predict surgeon performance across distinct data samples. Similar to reliability, we characterized agreement of $\geq 80\%$ as good and $\geq 90\%$ as excellent. For comparison, we also calculated agreement based on an approximation of the ProPublica Surgeon Scorecard methodology using MSQC data.⁷ All analyses were performed using SAS version 9.4 software (SAS Institute, Inc., Cary, NC).

Results

Structural, care process, and patient characteristics from the 7032 cases included in the scorecard are summarized in Table 3. We generated composite scores utilizing the 5 domain criteria to

FIGURE 1. ONLINE ONLY—Distribution of Scores for the Colectomy Scorecard

TABLE 3. Structural, Care Process, and Patient Characteristics by Surgeon Star Rating for the Colectomy Scorecard

Characteristic	Surgeon Quality Star Rating			Total
	1-Star (low outlier)	2-Star (middle tier)	3-Star (high outlier)	
Hospitals N	22	63	29	64
Surgeons N	18	415	36	469
Patients N	665	4655	1712	7032
STRUCTURAL				
Beds, mean (sd)	424 (210)	401 (237)	480 (281)	422 (248)
Academic Medical Center, n (%)	307 (46.2)	820 (17.6)	301 (17.6)	1428 (20.3)
Teaching, n (%)	285 (42.9)	2890 (62.1)	1203 (70.3)	4378 (62.3)
Statistical Area—Large, n (%)	567 (85.3)	3480 (74.8)	1147 (67)	5194 (73.9)
Statistical Area—Medium, n (%)	85 (12.8)	756 (16.2)	315 (18.4)	1156 (16.4)
Statistical Area—Noncore, Micro or Small, n (%)	13 (1.95)	419 (9)	250 (14.6)	682 (9.7)
[A] Survival				
Mortality, n (%)	5 (0.75)	53 (1.14)	7 (0.41)	65 (0.92)
[B] Morbidity Avoidance				
Overall Morbidity, n (%)	178 (26.8)	805 (17.3)	166 (9.7)	1149 (16.3)
Major Morbidity, n (%)	101 (15.2)	472 (10.1)	105 (6.13)	678 (9.64)
Sepsis—Any, n (%)	51 (7.67)	215 (4.62)	31 (1.81)	297 (4.22)
Surgical Site Infection—Total, n (%)	81 (12.2)	385 (8.27)	56 (3.27)	522 (7.42)
Surgical Site Infection—Superficial, n (%)	42 (6.32)	200 (4.3)	26 (1.52)	268 (3.81)
Surgical Site Infection—Deep, n (%)	12 (1.8)	57 (1.22)	10 (0.58)	79 (1.12)
Surgical Site Infection—Organ Space, n (%)	30 (4.51)	139 (2.99)	22 (1.29)	191 (2.72)
Clostridium Difficile, n (%)	17 (2.56)	43 (0.92)	7 (0.41)	67 (0.95)
[C] Anastomotic Leak Avoidance				
Anastomotic leak, n (%)	11 (1.65)	68 (1.46)	11 (0.64)	90 (1.28)

Characteristic	Surgeon Quality Star Rating			Total
	1-Star (low outlier)	2-Star (middle tier)	3-Star (high outlier)	
[D] SSI Bundle Compliance				
Total Score, mean (sd)	2.92 (0.96)	3.45 (0.97)	4.2 (0.98)	3.58 (1.04)
Appropriate IV Prophylactic Antibiotics, n (%)	304 (45.7)	2642 (56.8)	1071 (62.6)	4017 (57.1)
Postoperative Normothermia (temp>98.6°F), n (%)	590 (88.7)	4198 (90.2)	1570 (91.7)	6358 (90.4)
Oral Antibiotics with Mechanical Bowel Prep, n (%)	153 (23)	1161 (24.9)	851 (49.7)	2165 (30.8)
Postoperative Day-1 Glucose <=140mg/dL, n (%)	438 (65.9)	3210 (69)	1325 (77.4)	4973 (70.7)
Minimally Invasive Surgery, n (%)	295 (44.4)	2629 (56.5)	1462 (85.4)	4386 (62.4)
Short Operative Duration, n (%)	72 (10.8)	1151 (24.7)	464 (27.1)	1687 (24)
[*] Care Processes Other				
Antibiotics: Weight Adjusted Dose, n (%)	134 (20.2)	1274 (27.4)	523 (30.5)	1931 (27.5)
Antibiotics: Redosing, n (%)	19 (2.86)	109 (2.34)	81 (4.73)	209 (2.97)
Robotic Surgery, n (%)	131 (19.7)	406 (8.72)	118 (6.89)	655 (9.31)
Perioperative VTE Prophylaxis, n (%)	390 (58.6)	1700 (36.5)	796 (46.5)	2886 (41)
Postoperative VTE Prophylaxis, n (%)	601 (90.4)	4086 (87.8)	1507 (88)	6194 (88.1)
[E] Utilization				
Hospital Length of Stay (LOS), mean (sd)	6.98 (5.14)	6.24 (4.83)	4.97 (3.66)	6 (4.64)
Extended Hospital LOS, n (%)	55 (8.27)	284 (6.1)	51 (2.98)	390 (5.55)
ED visit, n (%)	82 (12.3)	419 (9)	129 (7.54)	630 (8.96)
Unplanned Readmission, n (%)	111 (16.7)	495 (10.6)	106 (6.19)	712 (10.1)
Unplanned Reoperation, n (%)	47 (7.07)	281 (6.04)	61 (3.56)	389 (5.53)

Note: Outcomes unadjusted for morbidity and mortality from 7,032 patients across colectomy procedures (see Methods inclusion criteria).

assess surgical performance in each procedure. Figure 1—*online only*—contains the total distribution of scores, which (with statistical confirmation) represents a normal distribution that is useful for identifying outliers in surgical performance. Composite scores were then attributed to the 469 surgeons that performed the surgeries. Figure 2 displays composite scores for all surgeons, with red and green lines representing statistically significant low- and high-performing outliers, respectively. Under our methodology, these statistical outliers represent clinically meaningful variation between low- and high-performing surgeons.

Figure 3 displays reliability curves for the overall composite score and each individual domain.

Reliability represents the ability of the model to discriminate between surgeons, despite measurement error. The x-axis shows the number of cases sampled per surgeon, whereas the y-axis indicates reliability (λ) from 0 to 1. Reliability is classified to be good for $\lambda \geq 0.8$ and excellent for $\lambda \geq 0.9$, represented by horizontal reference lines in Figure 3. The intersections of these reference lines with reliability curves indicate the minimum sample size for good and excellent reliability; for the composite score (dark, bolded line), this corresponds to 62 cases per surgeon for good reliability and 170 cases per surgeon for excellent reliability.

The predicted reliability curves in Figure 3 vary widely by domain; SSI bundle compliance has the

FIGURE 2. Rankings, Variation and Outlier Discrimination for the Colectomy Scorecard, MSQC 2012–2015

*This measure scores 1 point for each of the five domains: survival, morbidity avoidance, anastomotic leak avoidance, SSI colectomy bundle, and utilization.

**Star rating is based on statistical significance indicated by confidence intervals that do not overlap the overall average.

***CPT codes included are 44140, 44141, 44143, 44144, 44145, 44146, 44147, 44150, 44151, 44155, 44156, 44157, 44158, 44160, 44204, 44205, 44206, 44207, 44208, 44210, 44211, and 44212.

highest reliability, followed by morbidity avoidance and utilization. The clustering of these domains with the composite score indicates that the relative domain weights achieve a good balance in assessing care processes, morbidity, and utilization in the composite score. Although SSI bundle compliance actually displays slightly higher reliability than the composite score, it is important to note that the other domains also made meaningful, independent contributions to reliability, even at smaller sample sizes. These domains converge on excellent reliability with increasing sample size.

Anastomotic leak avoidance and survival have the lowest reliability. Both mortality and anastomotic leak are rare events that tend not to cluster by surgeon; it is thus infeasible to attain a sufficient

sample size for good reliability. This ordering of domain reliabilities is consistent with conceptual expectations of care processes closely linked to surgeons and complications strongly influenced by non-modifiable factors.

Finally, we calculated the test-retest agreement (proportion of surgeons with identical star ratings on repeated measurement) for our scorecard methodology to be 89.0% ($p < 0.01$, 95% CI 88.3–93.2). This indicates a nearly excellent level of repeatability with our methodology when surgeon star ratings are compared across distinct data samples. By comparison, agreement for ProPublica's Surgeon Scorecard was found to be 32.3% ($p < 0.01$, 95% CI 15.3–43.9), which is consistent with reports by Ban et al and Auffenberg et al.^{31,32}

FIGURE 3. Reliability vs. Sample Size for the Composite Scorecard and Individual Domains

Reliability (λ) as a function of sample size (n) is plotted for the composite measure, as well as for each of the 5 domains of the composite measure: [A] survival, [B] morbidity avoidance, [C] anastomotic leak avoidance, [D] SSI bundle compliance, and [E] utilization. Horizontal lines at $\lambda = 0.8$ and $\lambda = 0.9$ indicate good and excellent reliability, respectively.

This difference in test-retest agreement is depicted in Figure 4, with each “circle” in a Venn diagram representing the score of surgeons in a single measurement or time period and the overlap representing test-retest agreement.

Discussion

In publishing their version of a surgeon scorecard, ProPublica underscored the need for public scrutiny on individual surgeon performance. No one can argue that the general public should not have some reliable means to assess a surgeon’s skill and expertise before embarking on a potentially lifesaving surgical procedure. However, although this information is valuable, the methodology behind it must be statistically valid and reliable. In this article, we present a scorecard for colectomy surgeries with a multidimensional approach using both clinical and administrative domains in an attempt to strengthen methodology and predict future performance.

Previous efforts to develop a measure assessing surgical skill have fallen short largely because the measurements do not convey a comprehensive or cohesive account of the entire surgical episode.^{8,10,11} For example, ProPublica’s Surgeon Scorecard utilized only administrative billing data

limited to Medicare patients, leading to inadequate sample size for reliable measures.²⁰ It relied primarily on mortality and an “adjusted complication rate” that encompasses only the 30-day readmission rate.⁸ However, 88% of complications occur during the index hospitalization; this data is important, if not essential, to capture in a scorecard.^{9,10} Additionally, their scorecard uses a “hypothetical average hospital,” thus failing to account for possible variation in quality of surgeons between high- and low-performing hospitals.¹⁰ Our scorecard attempts to address some of these shortfalls by utilizing clinical data from colectomy procedures to create a ranking based on survival, morbidity avoidance, anastomotic leak avoidance, SSI bundle compliance, and utilization.

Finally, an important characteristic of our scorecard is the ability to reproduce results for future performance with test-retest agreement for 1-star and 3-star surgeons evaluated at 2 distinct time points. We calculated agreement for ProPublica’s Surgeon Scorecard to be 32.3%. Conversely, agreement for the Colectomy Scorecard was 89%, suggesting it has the potential to more reliably predict future surgeon performance. With increasingly robust measures of surgical quality, hospitals can more efficiently and effectively design targeted

FIGURE 4. Test-Retest Agreement for ProPublica’s Surgeon Scorecard vs. our Colectomy Scorecard Agreement, or the proportion of surgeons with identical outlier star ratings on repeated measurement, is depicted here for the ProPublica Surgeon Scorecard and our Colectomy Scorecard for 1-star and 3-star surgeons.

interventions to improve surgical performance, promote transparency between providers and patients, and ultimately improve patient care.

Our study has several limitations that must be addressed. First, the data from our scorecard is from the MSQC; these data may not accurately represent a national cohort, and such high-quality data is not available in other regions. Second, the accuracy of the registry data is limited, especially in the domain of case-mix complexity. We utilized CPT codes to identify specific colectomy procedures, but it is possible that some cases were misclassified or case complexity is inadequately measured. Third, risk adjustment is inherently imperfect and may have confounded our study in some immeasurable way. Future scorecards could potentially enhance the risk adjustment to evaluate if it markedly improves the reliability signal. Fourth, our scorecard only included data for surgeons that performed colectomies. Although the underlying methodology of our scorecard may translate to other procedures, the domains and items utilized in our scorecard are specific to colectomies. Scorecards for other procedures would likely require modification to these elements and weights with subsequent testing for statistical agreement and reliability in predicting future performance. This does limit the scope of our Colectomy Scorecard, but the intent of this article is primarily to illustrate a proof of concept with wide applicability across surgical specialties. We strongly advocate for the construction of additional scorecards with novel methodological improvements for different procedures. Finally, the 5 criteria by which surgeons are evaluated were selected by a limited number of MSQC surgeons. Although these measures reliably predicted surgeon performance, future scorecards should investigate whether other measures provide superior reliability when included in the composite score, thus using more empiric approaches to measure selection.

A major goal for any publicly reported scorecard should be to provide a pathway to improvement for surgeons. With the addition of robust

measures to improve reliability, scorecards can be increasingly used by hospitals and surgical leaders to identify and target high and low performance. Our scorecard provides domain-specific feedback, allowing surgeons to identify specific areas in which they might improve care in the operating room and, more broadly, during the patient's surgical journey. Furthermore, these data can be used within the surgical community to develop targeted interventions to assist low-performing surgeons. For example, hospitals could institute mentorship programs by identifying high- and low-performing surgeons within a particular scorecard domain and pairing them together with the hope of increasing the low-performing surgeon's performance. Scorecard reports could also be combined with other validated assessment criteria, such as a videotaped review of surgeries, to provide a more robust evaluation of individual surgeons.¹³ With reliable and reproducible measures in place, the scorecard can make meaningful and impactful steps toward improving surgical proficiency and outcomes.

Although controversial, the inclusion of patient-reported quality measures for a scorecard deserves mention. It is debatable whether or not patient-reported ratings of surgeon skill and quality are aligned with a surgeon's technical skill and also which measures are the most valid and reliable.³³ Although we did not include patient-reported ratings in our methodology, adding a patient-driven focus domain deserves further investigation and research.

Conclusion

The ability to evaluate surgeons, the episode of surgery, and the quality of outcomes is becoming a mainstay in healthcare. It is essential to continue strengthening existing methodologies to develop the next generation of scorecards. At the University of Michigan, we have developed the Colectomy Scorecard, a report card that provides a composite score for individual surgeons performing colectomies based on MSQC clinical data. Through validation testing, this scorecard was found to

reliably predict future surgeon performance. By providing domain-specific feedback to individual surgeons, our scorecard attempts to introduce a methodology that could help surgeons and hospitals develop targeted interventions to improve outcomes. Reliable surgeon scorecards are a step in the right direction to improving transparency, value, and quality of healthcare outcomes.

References

1. Mongan JJ, Ferris TG, Lee TH. Options for slowing the growth of health care costs. *N Engl J Med*. 2008;358(14):1509–1514. doi:10.1056/NEJMs0707912
2. Blendon RJ, Benson JM, Hero JO. Public trust in physicians—U.S. medicine in international perspective. *N Engl J Med*. 2014;371(17):1570–1572. doi:10.1056/NEJMp1407373
3. Levey NN. Medical professionalism and the future of public trust in physicians. *JAMA*. 2015;313(18):1827–1828. doi:10.1001/jama.2015.4172
4. Shea KK, Shih A, Davis K, Commonwealth Fund. *Health Care Opinion Leaders' Views on the Transparency of Health Care Quality and Price Information in the United States*. <https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101462076-pdf>. November 2007.
5. Fung CH, Lim Y-W, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Ann Intern Med*. 2008;148(2):111–123. doi: 10.7326/0003-4819-148-2-200801150-00006
6. Surgeon Scorecard. ProPublica; 2015. <https://projects.ProPublica.org/surgeons/>. Accessed September 11, 2015.
7. Pierce O, Allen M. Assessing surgeon-level risk of patient harm during elective surgery for public reporting (as of August 4, 2015). <https://static.propublica.org/projects/patient-safety/methodology/surgeon-level-risk-methodology.pdf>. Accessed August 4, 2015.
8. Friedberg MW, Pronovost PP, Shahian DM, et al. *A Methodological Critique of the ProPublica Surgeon Scorecard*. Santa Monica, CA: RAND Corporation; 2015.
9. Agency for Healthcare Research and Quality. *Patient Safety Quality Indicators Composite Measure Workgroup: Final Report*. Rockville, MD: AHRQ Quality Indicators; 2008. https://qualityindicators.ahrq.gov/Downloads/Modules/PSI/PSI_Composite_Development.pdf.
10. Friedberg MW, Bilimoria KY, Pronovost PJ, Shahian DM, Damberg CL, Zaslavsky AM. *Response to ProPublica's Rebuttal of Our Critique of the Surgeon Scorecard*. Santa Monica, CA: RAND Corporation; 2015.
11. Sekhri S, Mossner J, Iyengar R, Mullard A, Englesbe MJ, Papin J. Evaluating surgeon scorecards. *Mich J Med* 2016;1(1):74–77. doi:10.3998/mjm.13761231.0001.113
12. Campbell DA Jr, Kubus JJ, Henke PK, Hutton M, Englesbe MJ. The Michigan Surgical Quality Collaborative: a legacy of Shukri Khuri. *Am J Surg*. 2009;198(5 Suppl):S49–S55. doi:10.1016/j.amjsurg.2009.08.002
13. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(15):1434–1442. doi:10.1056/NEJMsa1300625
14. Campbell DA Jr, Englesbe MJ, Kubus JJ, et al. Accelerating the pace of surgical quality improvement: The power of hospital collaboration. *Arch Surg-Chicago*. 2010;145(10):985–991. doi:10.1001/archsurg.2010.220
15. Murray ACA, Chiuzaan C, Kiran RP. Risk of anastomotic leak after laparoscopic versus open colectomy. *Surg Endosc*. 2016;30(12):5275–5282. doi:10.1007/s00464-016-4875-0
16. Jaffe TA, Meka AP, Semaan DZ, et al. Optimizing value of colon surgery in Michigan. *Ann Surg*. 2017;265(6):1178–1182. doi:10.1097/SLA.0000000000001880
17. Waits S, Fritze D, Banerjee M, et al. Developing an argument for bundled interventions to reduce surgical site infection in colorectal surgery. *Surgery*. 2014;155(4):602–606. doi:10.1016/j.surg.2013.12.004
18. Agresti A. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley; 2013.
19. Hox JJ. *Multilevel Analysis: Techniques and Applications*. Lawrence Erlbaum Associates; 2002.
20. Jaffe TA, Hasday SJ, Dimick JB. Power outage— inadequate surgeon performance measures leave patients in the dark. *JAMA Surg*. 2016;151(7):599–600. doi:10.1001/jamasurg.2015.5459
21. Blom G. *Statistical Estimates and Transformed Beta-Variables*. New York, NY: Wiley; 1958.
22. Cochran WG. *Planning and Analysis of Observational Studies*. New York, NY: Wiley; 2009.
23. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. *N Engl J Med*. 1984;311(7):442–448. doi:10.1056/NEJM198408163110705

24. Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the analyses of variance and covariance. *Rev Educ Res.* 1972;42(3):237–288. doi:10.3102/00346543042003237
25. Snijders TAB, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London, UK: Sage Publications; 2012.
26. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage Publications; 2002.
27. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol.* 2006;59(10):1033–1039. doi:10.1016/j.jclinepi.2005.10.015
28. Kottner J, Streiner DL. The difference between reliability and agreement. *J Clin Epidemiol.* 2011;64(6):701–702. doi:10.1016/j.jclinepi.2010.12.001
29. Donner A, Wells G. A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics.* 1986;42(2):401–412.
30. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res.* 2004;13(4): 251–271. doi:10.1191/0962280204sm365ra
31. Ban KA, Cohen ME, Ko CY, et al. Evaluation of the ProPublica Surgeon Scorecard “adjusted complication rate” measure specifications. *Ann Surg.* 2016;264(4):1. http://www.uphs.upenn.edu/surgery/Education/PEBLR/Aug_2016/ProPublica_AnnSurg2016.pdf.
32. Auffenberg GB, Ghani KR, Ye Z, et al. Comparison of publicly reported surgical outcomes with quality measures from a statewide improvement collaborative. *JAMA Surg.* 2016;151(7):680–682. doi:10.1001/jamasurg.2016.0077
33. Cohen JB, Myckatyn TM, Brandt K. The importance of patient satisfaction: a blessing, a curse, or simply irrelevant? *Plast Reconstr Surg.* 2017;139(1):257–261. doi:10.1097/PRS.0000000000002848