Evaluation of the ProPublica Surgeon Scorecard ''Adjusted Complication Rate'' Measure Specifications

Kristen A. Ban, MD,† Mark E. Cohen, PhD,* Clifford Y. Ko, MD,*‡ Mark W. Friedberg, MD,§¶ Jonah J. Stulberg, MD,|| Lynn Zhou, PhD,* Bruce L. Hall, MD,**††‡‡§§ David B. Hoyt, MD,* and Karl Y. Bilimoria, MD, MS*||*

Objectives: The ProPublica Surgeon Scorecard is the first nationwide, multispecialty public reporting of individual surgeon outcomes. However, ProPublica's use of a previously undescribed outcome measure (composite of in-hospital mortality or 30-day related readmission) and inclusion of only inpatients have been questioned. Our objectives were to (1) determine the proportion of cases excluded by ProPublica's specifications, (2) assess the proportion of inpatient complications excluded from ProPublica's measure, and (3) examine the validity of ProPublica's outcome measure by comparing performance on the measure to well-established postoperative outcome measures.

Methods: Using ACS-NSQIP data (2012–2014) for 8 ProPublica procedures and for All Operations, the proportion of cases meeting all ProPublica inclusion criteria was determined. We assessed the proportion of complications occurring inpatient, and thus not considered by ProPublica's measure. Finally, we compared risk-adjusted performance based on ProPublica's measure specifications to established ACS-NSQIP outcome measure performance (eg, death/serious morbidity, mortality).

Results: ProPublica's inclusion criteria resulted in elimination of 82% of all operations from assessment (range: 42% for total knee arthroplasty to 96% for laparoscopic cholecystectomy). For all ProPublica operations combined, 84% of complications occur during inpatient hospitalization (range: 61% for TURP to 88% for total hip arthroplasty), and are thus missed by the ProPublica measure. Hospital-level performance on the ProPublica measure correlated weakly with established complication measures, but correlated strongly with readmission ($R^2 = 0.834$, $P < 0.001$).

Conclusions: ProPublica's outcome measure specifications exclude 82% of cases, miss 84% of postoperative complications, and correlate poorly with well-established postoperative outcomes. Thus, the validity of the ProPublica Surgeon Scorecard is questionable.

Keywords: ProPublica, public reporting, Surgeon Scorecard, surgeon-specific reporting

(*Ann Surg* 2016;xx:xxx–xxx)

Public reporting of individual provider performance has the potential to improve transparency for patients and to drive quality improvement in surgical care. Public reporting of performance at the hospital-level is widespread, but until recently, large-scale public reporting of individual surgeon performance was not available. Developed by journalists, the ProPublica Surgeon Scorecard is the first nationwide, multispecialty public reporting of individual surgeon outcomes.[1] The Surgeon Scorecard reports ''Adjusted Complication Rates'' for individual surgeons performing 1 of 8 elective inpatient surgical procedures using a previously undescribed outcome measure, a composite of inpatient mortality, or 30-day related readmission.[2,3]

Following public release of the Surgeon Scorecard, many methodologists raised concerns about ProPublica's use of a previously undescribed ''Adjusted Complication Rate'' measure as a basis for public reporting, before testing or validating this new measure.[3]Additional concerns included problems with the source data (including misattribution of surgical cases to the wrong physicians), inadequate case-mix adjustment, unknown and apparently poor measurement reliability, and the exclusion of complications during the index admission (other than death)—a notable departure from any well-accepted, previously described postoperative outcome measure of complications.[3]

In addition, the Surgeon Scorecard was limited to examining inpatient surgical procedures. This is particularly concerning since most of the procedures within the Surgeon Scorecard are typically, or even predominantly, performed as outpatient or short-stay procedures. The limitation of the Surgeon Scorecard to inpatient encounters might result in disproportionate capture of complex cases (ie, laparoscopic cholecystectomy with length of stay of 2 days or more). Moreover, this limitation could also severely decrease the

From the *American College of Surgeons, Chicago, IL; †Department of Surgery, Loyola University Medical Center, Maywood, IL; ‡Department of Surgery, University of California Los Angeles, Los Angeles, CA; §RAND Corporation, Boston, MA; ¶Brigham and Women's Hospital, Harvard Medical School, Boston, MA; ||Surgical Outcomes and Quality Improvement Center, Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, IL; **Department of Surgery, Washington University in St. Louis, St. Louis, MO; ††Center for Health Policy and the Olin Business School at Washington University in St. Louis, St. Louis, MO; ‡‡John Cochran Veterans Affairs Medical Center, St. Louis, MO; and §§BJC Healthcare, St. Louis, MO.

Reprints: Karl Y. Bilimoria, MD, MS, Division of Research and Optimal Patient Care, American College of Surgeons and Feinberg School of Medicine, Northwestern University, 633 N St Clair St, 22nd Floor, Chicago, IL 60611. E-mail: kbilimoria@facs.org.

Presented at the American Surgical Association, Chicago, IL, April 14–16, 2016.

KAB—involved in study conceptualization and design. Carried out the bulk of study analysis. Wrote substantial portion of the manuscript. Gave final approval prior to submission.

MEC—involved in study design, oversaw statistical analyses, proofread analytic code. Contributed to portions of the manuscript relating to statistical analysis. Gave final approval prior to submission.

CYK—involved in study conceptualization, provided feedback on initial drafts on the manuscript. Gave final approval prior to submission.

MWF—involved in study conceptualization and design, contributed substantially to the final manuscript. Gave final approval prior to submission.

JJS—involved in study conceptualization and ongoing design, contributed substantially to the manuscript. Gave final approval prior to submission.

LZ—wrote significant portions of code used in statistical analysis. Gave final approval prior to submission.

BLH—involved in study conceptualization, provided feedback on initial drafts on the manuscript. Gave final approval prior to submission.

DBH—provided feedback during and after completion of study analysis. Gave final approval prior to submission.

KYB—involved in study conceptualization and design, ongoing involvement in data analysis. Contributed substantially to manuscript. Gave final approval prior to submission.

AHRQ R21 (KYB): ''Engaging Patients and Providers to Expand Public Reporting in Surgery.''

The authors report no conflicts of interest.

ISSN: 0003-4932/14/26105-0821

DOI: 10.1097/SLA.0000000000001858

number of cases eligible for modeling, thus decreasing the reliability of those estimates.[4,5]

The objectives of this study were to use high-quality clinical registry data (1) to determine what proportion of cases were excluded by ProPublica's specifications (eg, inpatient, nonemergent, age ≥65), (2) to assess the proportion of complications occurring as an inpatient, and thus not considered in ProPublica's measure, and (3) to examine the validity of ProPublica's outcome measure by comparing performance on ProPublica's outcome measure to well-established and previously validated postoperative outcome measures.

METHODS

Data Source and Study Population

Data were obtained for all operations performed between January 1, 2012 and December 31, 2014 included in the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP). The details of ACS NSQIP have been previously described; but in brief, clinical data are collected in a standardized format by trained, certified, and audited abstractors at each hospital.[6] Thirty-day postoperative outcomes are ascertained from the medical records, discussions with providers, and via direct communication with patients when needed. Outcomes are ascertained irrespective of whether the patient is an inpatient, discharged home, or transferred/readmitted to an outside facility. Variables collected include patient demographics, preoperative risk factors, laboratory values, operative variables, readmission to any hospital, and numerous postoperative complications.

Our study population was defined by the inclusion and exclusion criteria employed by ProPublica in the development of their Surgeon Scorecard.[2] Namely, all patients younger than 65 years of age and those who did not have an inpatient admission (defined as a length of stay greater than 24 hours) were excluded. Emergency cases were also excluded. ProPublica limited their surgeon-level analyses to surgeons performing 20 or more cases. We did not limit our analyses based on hospital case volume since such restrictions were not necessary to compare hospital-level performance on the ProPublica measure to standard postoperative outcomes or to determine the proportion of complications occurring as an inpatient.

The ProPublica Surgeon Scorecard examined 8 individual procedures: laparoscopic cholecystectomy, radical prostatectomy, transurethral prostatectomy (TURP), cervical fusion of the anterior column (anterior technique), lumbar and lumbosacral fusion of the anterior column, posterior technique (ALIF), lumbar and lumbosacral fusion of the posterior column, posterior technique (PLIF), total hip arthroplasty, and total knee arthroplasty, all based on ICD-9 procedure codes. ICD-9 procedures codes are not used in ACS NSQIP, so we used the corresponding Current Procedural Terminology codes to identify the individual surgeries listed above. Bilateral total hip and knee arthroplasty surgeries were excluded (revisional surgeries could not be excluded). Although not done by ProPublica, we also analyzed "All Operations" in the ACS NSQIP dataset. This allowed us to compare the correlation between ProPublica's outcome measure and well-established postoperative outcomes using a much larger sample size than would be available for each individual ProPublica procedure.

Outcomes

We assessed the number of cases excluded by ProPublica's specifications in the All Operations group and in each of the 8 ProPublica procedures. To determine the number of cases excluded by ProPublica's specifications, each ProPublica inclusion criterion was applied sequentially to the cohort of interest. These inclusion criteria in order applied were inpatient status, nonemergent surgery, and age 65 and older. After application of each successive criterion, the number of cases eligible for inclusion was recalculated.

Standard postoperative outcomes evaluated included mortality, death or serious morbidity (DSM, morbidities defined by NSQIP), any morbidity, readmission within 30 days of index operation, and surgical site infection (SSI), which is the most common postoperative complication overall. As defined by ACS NSQIP in each of their semiannual reports,[7] serious morbidity includes cardiac arrest or myocardial infarction, deep venous thrombosis, pulmonary embolism, sepsis or septic shock, deep wound infection, organ space infection, dehiscence, reintubation, pneumonia, renal failure or insufficiency, urinary tract infection, or reoperation. Any morbidity includes serious morbidity plus bleeding events, superficial SSI, stroke or peripheral neurological complication, or graft failure.

The percentage of complications occurring as an inpatient was determined for all 8 of the ProPublica procedures, as well as the combination of all ProPublica procedures together. Complications included in the definition of "any morbidity" were evaluated in this analysis, with the exception of return to the OR, which was not included. We determined whether a complication occurred as an inpatient or as an outpatient by comparing the date of the complication to the date of discharge. Complications occurring on or prior to the date of discharge were classified as inpatient, whereas complications occurring after the date of discharge were classified as outpatient.

Using ACS NSQIP clinical data, we reconstructed the ProPublica "adjusted complication rate" outcome measure, which was originally created using administrative claims data. We recreated a composite of inpatient mortality or unplanned readmission for a related complication within 30 days of the index surgery. A readmission was considered to be a related complication if it fell into 1 of 3 categories: (1) the reason for the readmission was coded as a standard ACS NSQIP postoperative complication (the clinical data abstractor has standardized definitions to determine whether or not this was the case); (2) the readmission was not coded as a standard ACS NSQIP postoperative complication, but had an admission diagnosis ICD-9 code considered related to the surgery based on ICD-9 codes as defined by Merkow et al[8] (ie, when the abstractor does not believe the reason for readmission falls under a standard ACS NSQIP postoperative complication, they must enter an ICD-9 code as the reason for the readmission); (3) the readmission was coded using 1 of certain additional ICD-9 codes listed in the ProPublica appendices specific to the procedure in question.[9] Hospitals had the option to record ICD-10 codes as of 2014. ICD-10 codes corresponding to the ICD-9 codes described above were also included.

Correlations between hospital-level performance on ProPublica's outcome measure and these standard NSQIP outcome measures were determined in the All Operations group and each of the 8 ProPublica procedure groups.

Statistical Analysis

Modeling of both the ProPublica outcome measure and typical postoperative outcomes was carried out according to the statistical methodology employed by ACS NSQIP, which has been described previously.[5] In this method, hierarchical logistic regression considers patients nested within hospitals, with hospital as a random effect. Hierarchical modeling incorporates both risk adjustment and an empirical Bayes-type shrinkage adjustment in the estimated odds ratio. For all outcomes, a logistic model and forward selection were used to identify a set of predictive variables for use in the hierarchical model. These variables were selected from patient demographics, patient comorbidities, health summary status variables [eg,

functional status, American Society of anesthesiologists (ASA) class], and procedure variables. Correlation in odds ratios between the ProPublica outcome and each of the typical ACS NSQIP postoperative outcomes was evaluated using R^2 values at the hospital-level. Analyses were performed at the hospital-level as there were insufficient case numbers at the surgeon-level to run the models and obtain reliable estimates. Moreover, our objective was to assess the specifications of the novel ProPublica measure, not to recreate the surgeon-level assessment.

All analyses were performed using SAS version 9.4 (SAS Institute, Chicago, IL).

RESULTS

The All Operations cohort included 353,015 patients from 550 hospitals. Median age for All Operations was 73 (68–79 IQR). Additional patient characteristics by procedure are detailed in Table 1.

In the All Operations group and across the 8 individual ProPublica procedures examined, a considerable proportion of cases were excluded by ProPublica's inclusion criteria (inpatient status, nonemergent, age ≥65). The percentage of patients *excluded* from analysis after application of all criteria was 82% for All Operations, and ranged from 42% for total knee arthroplasty to as high as 96% for laparoscopic cholecystectomy (Table 2).

Event rates in the cohort meeting all ProPublica inclusion criteria were considerably higher compared with event rates for all patients undergoing the same procedure without inpatient, nonemergent, and age restrictions (Table 3). In laparoscopic cholecystectomy, for example, the ProPublica ''adjusted complication rate'' in patients meeting all ProPublica inclusion criteria was 6.0%, compared with 1.7% in all patients. The mortality rate in cholecystectomy patients meeting all ProPublica criteria was 1.6% compared with 0.2% in all patients undergoing the procedure. In every patient cohort, the rates of NSQIP death or serious morbidity were higher than the rates produced by the ProPublica ''adjusted complication rate'' measure. For example, in all patients undergoing All Operations, the rate of death or serious morbidity was 7.8% compared with the ProPublica event rate of 3.3%.

Analysis of data for each of the 8 procedures in the ProPublica Surgeon Scorecard demonstrated that the majority of complications (84%) occur as an inpatient and are therefore not included inProPublica's ''adjusted complication rate'' outcome measure (Table 4). The proportion of complications occurring as an inpatient ranged from 61% for TURP to 88% for ALIF and total hip arthroplasty.

Hospital-level performance scores calculated by applying the ProPublica ''adjusted complication rate'' measure specifications to NSQIP data correlated weakly with previously well-established surgical outcome measures including mortality, death/serious morbidity, any morbidity, and SSI (Table 5). For All Operations, correlation ranged from $R^2 = 0.077$ ($P < 0.001$) for any morbidity to $R^2 = 0.229$ ($P < 0.001$) for death/serious morbidity. Conversely, the ProPublica outcome measure correlated very strongly with readmission, $R^2 = 0.834$ ($P < 0.001$). These results were similar for each of the 8 ProPublica procedures; however, numerous models were unstable due to low event rates and/or failed to identify hospital random effects.

DISCUSSION

Using robust clinical data, we found that ProPublica's specifications (inpatient, nonemergency surgery, age ≥65) exclude the majority of cases performed and fail to account for the majority of postoperative complications, since they occur in the inpatient setting for the procedures evaluated in the Surgeon Scorecard. Further, this group represents a unique, high-risk subset of patients, as evidenced by outcome event rates 2 to 8.5 times higher than in all patients undergoing the same procedures. Hospital-level performance calculated using ProPublica's previously undescribed outcome measure correlates poorly with well-established postoperative outcomes with the exception of readmission.

Our analysis demonstrates that for each of the ProPublica procedures, ProPublica's specifications (inpatient, nonemergent, age ≥65) exclude a significant proportion of cases from their assessments. In the All Operations cohort, only 18% of patients meet all of ProPublica's inclusion criteria; for laparoscopic cholecystectomy, this number drops to 4%. One problem with significant case loss is that outcome reporting, whether at the hospital or surgeon level, requires large sample sizes for reliability.[10] Calculations by Jaffe et al require individual surgeon case volumes of over 600 laparoscopic cholecystectomies to correctly identify surgeons with 1.5 times the average complication rate 80% of the time. Even for extremely poor performing surgeons with 3 times the average complication rate, the number of cases required would be 50, well above the current ProPublica minimum.[11] Significant case loss in the Surgeon Scorecard, a tool that is marketed to all potential surgical patients, is also problematic because ProPublica's analysis was carried out on, and applies to, a very select group of higher-risk patients. Older patients who have an inpatient hospitalization following an elective surgery, particularly many of those chosen by ProPublica, represent an inherently higher-than-average risk cohort.

The extremely select group of patients included in ProPublica's analysis is further illustrated by the higher event rates seen in patients meeting all of ProPublica's inclusion criteria compared with all patients undergoing the same procedure without age or admission status limitations. Most event rates were twice as high in the ProPublica cohort compared with all patients undergoing the same surgery, but some event rates were considerably higher. Mortality rates for the ProPublica cohort were 8 times higher for laparoscopic cholecystectomy and 8.5 times higher for cervical fusion. Thus, the outcomes from an extremely small, select, and higher-than-average risk group of patients are being used to evaluate the performance of individual surgeons.

ProPublica's measure is described as an ''Adjusted Complication Rate,'' but it does not consider the majority of postoperative complications for all surgeries evaluated since 61% to 88% occur as an inpatient and are not captured by a related readmission. Even more complications are excluded from ProPublica's assessments since there are additional complications occurring after discharge, including death, that do not result in a readmission. As a result, ProPublica's outcome measure is essentially a measure of a provider's readmission rate. While readmission is a well-accepted postoperative outcome, it is not typically attributable to the surgeon alone. Prior literature has shown that the major drivers of readmission are patient risk and social factors.[12,13] Because so many complications in the procedures evaluated by ProPublica occur during the inpatient setting, and are thus excluded from ProPublica's measure, it is unlikely that ProPublica's Surgeon Scorecard presents valid information about individual surgeon ''adjusted complication rates.'' Moreover, by calling their measure an ''adjusted complication rate'' while dramatically understating the true risk of operative complications, ProPublica's Surgeon Scorecard misinforms patients who are considering the pros and cons of undergoing an elective procedure.

We found that hospital-level performance on ProPublica's novel outcome measure correlated poorly with standard ACS NSQIP postoperative outcomes including mortality, death/serious morbidity, any morbidity, and individual complications like surgical site infection, though ProPublica's measure did correlate with readmission. Several factors may contribute to the ProPublica outcome measure's

TABLE 1. Patient Characteristics

	All Operations	Laparoscopic Cholecystectomy	Radical Prostatectomy	TURP (Transurethral Prostate Resection)	Cervical Fusion of Anterior Column	ALIF (Lumbar Fusion of Anterior Column), Posterior Approach	PLIF (Lumbar Fusion of Posterior Column), Posterior Approach	Total Knee Arthroplasty	Total Hip Arthroplasty
Patients	353,015	3430	3891	2662	1398	1006	1178	51,193	21,193
Hospitals	550	467	245	246	215	152	167	354	351
Age; years, median (IQR)	73 (68–79)	76 (70–82)	69 (66–72)	76 (71–82)	70 (67–75)	71 (67–75)	71 (68–76)	72 (68–77)	73 (68–79)
Sex									
Male	160,466 (45.5%)	1627 (47.4%)	3891 (100%)	2662 (100%)	671 (48.0%)	430 (42.7%)	488 (41.4%)	18,694 (36.5%)	10,707 (38.0%)
Female	192,549 (54.5%)	1803 (52.6%)			727 (52.0%)	576 (57.3%)	690 (58.6%)	32,499 (63.5%)	17,486 (62.0%)
Race/ethnicity									
White	277,729 (78.7%)	2624 (76.5%)	2536 (65.2%)	1371 (51.5%)	1143 (81.8%)	905 (90.0%)	989 (84.0%)	41,038 (80.2%)	22,661 (80.4%)
Black	24,522 (7.0%)	329 (9.6%)	289 (7.4%)	117 (4.4%)	146 (10.4%)	39 (3.9%)	58 (4.9%)	2583 (5.1%)	1295 (4.6%)
Asian	9219 (2.6%)	156 (4.6%)	119 (3.1%)	99 (3.7%)	25 (1.8%)	17 (1.7%)	15 (1.3%)	1344 (2.6%)	473 (1.7%)
American Indian or Alaska Native	1279 (0.4%)	7 (0.2%)	5 (0.1%)	0 (0%)	11 (0.8%)	1 (0.1%)	3 (0.3%)	198 (0.4%)	72 (0.3%)
Native Hawaiian or Pacific Islander	938 (0.3%)	6 (0.2%)	12 (0.3%)	6 (0.2%)	8 (0.6%)	2 (0.2%)	3 (0.3%)	160 (0.3%)	46 (0.2%)
Unknown	39,328 (11.1%)	308 (9.0%)	930 (23.9%)	1069 (40.2%)	65 (4.7%)	42 (4.2%)	110 (9.3%)	5870 (11.5%)	3646 (12.9%)
Hispanic									
Yes	15,141 (4.3%)	284 (8.3%)	119 (3.1%)	142 (5.3%)	81 (5.8%)	41 (4.1%)	51 (4.3%)	2328 (4.6%)	612 (2.2%)
No	296,598 (84.0%)	2836 (82.7%)	2822 (72.5%)	1394 (52.4%)	1261 (90.2%)	929 (92.4%)	1024 (86.9%)	42,228 (82.5%)	23,363 (82.9%)
Unknown	41,276 (11.7%)	310 (9.0%)	950 (24.4%)	1126 (42.3%)	56 (4.0%)	36 (3.6%)	103 (8.7%)	6637 (13.0%)	4218 (15.0%)
BMI class									
Normal	93,624 (26.5%)	820 (23.9%)	781 (20.1%)	887 (33.3%)	287 (20.5%)	201 (20.0%)	228 (19.4%)	6394 (12.5%)	6523 (23.1%)
Underweight	8052 (2.3%)	62 (1.8%)	18 (0.5%)	37 (1.4%)	17 (1.2%)	11 (1.1%)	14 (1.2%)	131 (0.3%)	305 (1.1%)
Overweight	122,348 (34.7%)	1254 (36.6%)	1836 (47.2%)	1084 (40.7%)	492 (35.2%)	359 (35.7%)	454 (38.5%)	16,351 (31.9%)	10,358 (36.7%)
Class 1 Obese	75,750 (21.5%)	770 (22.5%)	960 (24.7%)	474 (17.8%)	363 (26.0%)	276 (27.4%)	291 (24.7%)	15,030 (29.4%)	6666 (23.6%)
Class 2 Obese	33,197 (9.4%)	320 (9.3%)	234 (6.0%)	132 (5.0%)	164 (11.7%)	121 (12.0%)	118 (10.0%)	8243 (16.1%)	2899 (10.3%)
Class 3 Obese	20,044 (5.7%)	204 (6.0%)	62 (1.6%)	48 (1.8%)	75 (5.4%)	38 (3.8%)	73 (6.2%)	5044 (9.9%)	1442 (5.1%)
Specialty									
General surgery	120,533 (34.1%)	3430 (100%)							
Orthopedics	120,477 (34.1%)				419 (30.0%)	499 (49.6%)	355 (30.1%)	51,193 (100%)	28,193 (100%)
Peripheral Vascular	41,788 (11.8%)								
Urology	20,828 (5.9%)		3891 (100%)	2662 (100%)					
Neurosurgery	18,851 (5.3%)				979 (70.0%)	507 (50.4%)	823 (69.9%)		
Thoracic surgery	11,331 (3.2%)								
Gynecology	9213 (2.6%)								
Cardiac surgery	5024 (1.4%)								
Otolaryngology	3234 (0.9%)								
Plastic surgery	1736 (0.5%)								

TABLE 2. Decrease in Number of Eligible Patients With Successive Application of ProPublica's Exclusion Criteria

	All Cases	Inpatient	Nonemergent	Age ≥65
Laparoscopic cholecystectomy				
N patients	83,264	12,782	8952	3430
% patients	100%	15.4%	10.8%	4.1%
Radical prostatectomy				
N patients	19,674	8545	8514	3891
% patients	100%	43.4%	43.3%	19.8%
TURP (transurethral prostate resection)				
N patients	11,468	3604	3248	2662
% patients	100%	31.4%	28.3%	23.2%
Cervical fusion of anterior column				
N patients	17,867	5476	5203	1398
% patients	100%	30.6%	29.1%	7.8%
ALIF (lumbar fusion of anterior column), posterior approach				
N patients	3888	3408	3378	1006
% patients	100%	87.7%	86.9%	25.9%
PLIF (lumbar fusion of posterior column), posterior approach				
N patients	3623	3118	3082	1178
% patients	100%	86.1%	85.1%	32.5%
Total knee arthroplasty				
N patients	88,693	85,871	85,477	51,193
% patients	100%	96.8%	96.4%	57.7%
Total hip arthroplasty				
N patients	57,284	52,597	51,703	28,193
% patients	100%	91.8%	90.3%	49.2%
All Operations				
N patients	1941,251	940,906	799,121	353,015
% patients	100%	48.5%	41.2%	18.2%

Each column shows the number of eligible cases after the application of that additional exclusion criterion. Successive exclusion criteria were applied moving from left to right, recalculating the number of eligible cases after each step.

failure to correlate with other postoperative outcome measures except readmission. Despite the inclusion of inpatient mortality in the ProPublica outcome measure, the total number of mortality events compared with readmissions is too small to contribute meaningfully to the measure. Outcomes including death/serious morbidity, any morbidity, and SSI capture complications that occur during both the inpatient and postdischarge outpatient phases of a patient's postoperative course. In this study, we found that the majority of complications following the specific surgeries evaluated by ProPublica occur as an inpatient prior to discharge. Because ProPublica's measure by construct fails to capture any complication that does not result in an inpatient death or a related readmission, it is not surprising that their measure correlates poorly with standard postoperative measures that do capture both inpatient and postdischarge complications.

In our analyses, performance on ProPublica's measure did not correlate well with most established postoperative outcomes, and it is likely that our methods inflate these correlations compared with what they would be using ProPublica's data at the surgeon-level. We have applied ProPublica's specifications and measure to high-quality clinical data using NSQIP's risk-adjustment methodology. The integrity of NSQIP data (correct case attributed to correct surgeon) and the case-mix adjustment methods employed in our models have been validated previously.[5] ProPublica's case-mix adjustment has not been validated, and the misattribution of cases to the wrong surgeon in the Surgeon Scorecard makes the integrity of ProPublica's data suspect.[3] Further, we are assessing correlations at the hospital-level with sample sizes that far exceed those used by ProPublica at the surgeon-level. By performing our analyses in this way, we increase the reliability of our estimates compared with ProPublica's analysis, which was performed on sample sizes as small as 20 cases. Although our analysis offers a best-case scenario for the ProPublica outcome measure, it still falls short and is likely not a useful measure of postoperative complications. Also concerning, many of our procedure-specific outcome models failed to converge, limited primarily by small event rates. These analyses were done at the hospital-level, where sample sizes are considerably higher compared with sample sizes at the surgeon level in ProPublica's analysis. These findings raise the question of how ProPublica was able to model individual surgeon outcomes using much smaller sample sizes when our hospital-level models failed.

There are several limitations of the current study. First, ProPublica used Medicare data, while we used ACS NSQIP data. These sources have different sampling methodology; Medicare captures all fee-for-service encounters based on administrative data, whereas NSQIP captures a systematic temporal sample of clinical data. We believe the use of clinical data from ACS NSQIP rather than administrative Medicare data may be particularly beneficial when attempting to understand postoperative complication data and assessing the ProPublica Scorecard outcome measure. ACS NSQIP hospitals are also a subset of the hospitals captured in the Medicare dataset and likely represent a selected group of hospitals with a focus on quality improvement.[14] Second, we chose not to perform our analyses at the individual surgeon level because of known challenges in assessing individual surgeon outcomes reliably,[10,15] and because sample sizes would have been extremely small at the individual surgeon level given NSQIP's sampling and the limitations imposed by the ProPublica exclusion criteria. Further, surgeon-level data was not necessary to assess the validity of the measure specifications themselves, which was our primary objective.

TABLE 3. ProPublica Measure and Well-established Postoperative Outcome Event Rates in All Patients (All Ages, Any Admission Status) Versus ProPublica Patients (Inpatients, Age ≥65)

	Cases	ProPublica Inpatient Death + Readmission Rate	Death Or Serious Morbidity Rate	Morbidity Rate	Mortality Rate	SSI Rate	Readmission Rate
All Operations (all patients, all ages)	1,941,251	63,007 (3.3%)	151,625 (7.8%)	246,488 (12.7%)	20,166 (1.0%)	52,115 (2.7%)	50,053 (2.6%)
All Operations (inpatient, age ≥65)	353,015	17,916 (5.1%)	47,254 (13.4%)	88,797 (25.2%)	6144 (1.7%)	14,223 (4.0%)	14,562 (4.1%)
Laparoscopic cholecystectomy (all patients, all ages)	83,264	1406 (1.7%)	2218 (2.7%)	2774 (3.3%)	161 (0.2%)	894 (1.1%)	1336 (1.6%)
Laparoscopic cholecystectomy (inpatient, age ≥65)	3430	204 (6.0%)	442 (12.9%)	506 (14.8%)	54 (1.6%)	101 (2.9%)	177 (5.2%)
Radical prostatectomy (all patients, all ages)	19,674	434 (2.2%)	976 (5.0%)	1738 (8.8%)	26 (0.1%)	241 (1.2%)	424 (2.2%)
Radical prostatectomy (inpatient, age ≥65)	3891	132 (3.4%)	342 (8.8%)	674 (17.3%)	12 (0.3%)	78 (2.0%)	128 (3.3%)
TURP (transurethral prostate resection) (all patients, all ages)	11,468	310 (2.7%)	829 (7.2%)	936 (8.2%)	70 (0.6%)	13 (0.1%)	282 (2.5%)
TURP (transurethral prostate resection) (inpatient, age ≥65)	2662	107 (4.0%)	299 (11.2%)	378 (14.2%)	35 (1.3%)	3 (0.1%)	96 (3.6%)
Cervical fusion of anterior column (all patients, all ages)	17,867	225 (1.3%)	591 (3.3%)	636 (3.6%)	43 (0.2%)	91 (0.5%)	201 (1.1%)
Cervical fusion of anterior column (inpatient, age ≥65)	1398	50 (3.6%)	173 (12.4%)	173 (12.4%)	23 (1.7%)	6 (0.4%)	36 (2.6%)
ALIF (lumbar fusion of anterior column), Posterior approach (all patients, all ages)	3888	95 (2.4%)	274 (7.1%)	628 (16.2%)	11 (0.3%)	71 (1.8%)	89 (2.3%)
ALIF (lumbar fusion of anterior column), Posterior approach (inpatient, age ≥65)	1006	34 (3.4%)	97 (9.6%)	250 (24.9%)	4 (0.4%)	24 (2.4%)	30 (3.0%)
PLIF (lumbar fusion of posterior column), Posterior approach (all patients, all ages)	3623	91 (2.5%)	242 (6.7%)	621 (17.1%)	9 (0.3%)	55 (1.5%)	87 (2.4%)
PLIF (lumbar fusion of posterior column), Posterior approach (inpatient, age ≥65)	1178	39 (3.3%)	104 (8.8%)	273 (23.2%)	5 (0.4%)	17 (1.4%)	36 (3.1%)
Total knee arthroplasty (all patients, all ages)	88,693	1506 (1.7%)	3680 (4.2%)	12,043 (13.6%)	100 (0.1%)	685 (0.8%)	1471 (1.7%)
Total knee arthroplasty (inpatient, age ≥65)	51,193	962 (1.9%)	2344 (4.6%)	7928 (15.5%)	69 (0.1%)	362 (0.7%)	939 (1.8%)
Total hip arthroplasty (all patients, all ages)	57,284	1059 (1.9%)	2527 (4.4%)	9251 (16.2%)	116 (0.2%)	671 (1.2%)	1000 (1.8%)
Total hip arthroplasty (inpatient, age ≥65)	28,193	598 (2.1%)	1465 (5.2%)	5552 (19.7%)	67 (0.2%)	304 (1.1%)	562 (2.0%)

TABLE 4. Proportion of Complications Occurring During Inpatient Stay (Not Included in ProPublica Surgeon Scorecard Outcome Measure)

	Cases	Any Complication Rate (n, %)		Inpatient Complication Rate (n, %)		Outpatient Complication Rate (n, %)		Proportion of Complications Occurring Inpatient (%)
Laparoscopic Cholecystectomy	3430	528	15.4%	349	10.2%	179	5.2%	66%
Radical Prostatectomy	3899	654	16.8%	471	12.1%	183	4.7%	72%
TURP (transurethral prostate resection)	2662	356	13.4%	218	8.2%	138	5.2%	61%
Cervical fusion of anterior column	1398	158	11.3%	101	7.2%	52	4.1%	64%
ALIF (lumbar fusion of anterior column), Posterior approach	1006	242	24.1%	214	21.3%	28	2.8%	88%
PLIF (lumbar fusion of posterior column), Posterior approach	1178	264	22.4%	226	19.2%	38	3.2%	86%
Total knee Arthroplasty	51,193	7818	15.3%	6659	13.0%	1159	2.3%	85%
total hip Arthroplasty	28,193	5392	19.1%	4754	16.9%	638	2.3%	88%
All ProPublica procedures combined	92,959	15,412	16.6%	12,992	14.0%	2415	2.6%	84%

TABLE 5. Correlation Between ProPublica Surgeon Scorecard Outcome Measure and Established Surgical Outcomes

	Correlation With ProPublica Measure				
	Death or Serious Morbidity	Any Morbidity	Mortality	SSI	Readmission
All Operations	$R^2 = 0.229$ ($P < 0.001$)	$R^2 = 0.077$ ($P < 0.001$)	$R^2 = 0.120$ ($P < 0.001$)	$R^2 = 0.125$ ($P < 0.001$)	$R^2 = 0.834$ ($P < 0.001$)
Laparoscopic Cholecystectomy	$R^2 = 0.188$ ($P < 0.001$)	*	$R^2 = 0.082$ ($P < 0.001$)	$R^2 = 0.094$ ($P < 0.001$)	$R^2 = 0.859$ ($P < 0.001$)
Radical Prostatectomy	$R^2 = 0.304$ ($P < 0.001$)	$R^2 = 0.177$ ($P < 0.001$)	*	$R^2 = 0.233$ ($P < 0.001$)	$R^2 = 0.972$ ($P < 0.001$)
TURP (transurethral prostate resection)	$R^2 = 0.176$ ($P < 0.001$)	$R^2 = 0.094$ ($P < 0.001$)	*	*	$R^2 = 0.910$ ($P < 0.001$)
Cervical fusion of anterior column	*	$R^2 = 0.194$ ($P < 0.001$)	*	*	*
ALIF (lumbar fusion of anterior column), posterior approach	*	$R^2 = 0.036$ ($P = 0.019$)	*	$R^2 = 0.233$ ($P < 0.001$)	$R^2 = 0.872$ ($P < 0.001$)
PLIF (lumbar fusion of posterior column), posterior approach	*	*	*	*	*
Total knee Arthroplasty	$R^2 = 0.177$ ($P < 0.001$)	$R^2 = 0.002$ ($P = 0.365$)	*	$R^2 = 0.092$ ($P < 0.001$)	$R^2 = 0.982$ ($P < 0.001$)
Total hip Arthroplasty	$R^2 = 0.229$ ($P < 0.001$)	$R^2 = 0.040$ ($P < 0.001$)	*	$R^2 = 0.139$ ($P < 0.001$)	$R^2 = 0.951$ ($P < 0.001$)

Assessed with Pearson correlation.
*Model failed to identify hospital random effects and/or too few events to model.

The stakes in the development of a new public reporting tool are high—misleading data are harmful to both patients and providers. The National Quality Forum outlines essential criteria that should be met with the introduction of any novel measure. Most important is the establishment of scientific acceptability, in other words, verification of the measure's reliability and validity.[16] The validity of ProPublica's previously undescribed measure was not determined prior to public release. Our study demonstrates that the ProPublica Surgeon Scorecard assessments are drawn from a sample that represents a narrow surgical population. Further, the ProPublica outcome measure captures only a small segment of postoperative complications. Although it is described as an ''adjusted complication rate,'' the ProPublica measure is a poor proxy for previously established measures of postoperative complications whose reliability and validity have been tested. All of these results suggest that the validity of ProPublica's measure and specifications are suspect. Both the public and the medical community agree that there is a need for increased transparency and public reporting for individual surgeons. The challenge remains in how to obtain adequate reliability with small sample sizes, particularly for less common surgeries. The development of reliable, valid, publicly reported surgeon-specific measures should remain a priority, but it will be crucial that these measures be properly evaluated prior to release considering the consequences for all parties involved.

REFERENCES

1. Wei S, Pierce O, Allen M. Surgeon Scorecard. 2015. Available at: https://projects.propublica.org/surgeons/. Accessed December 17, 2015.
2. Pierce O, Allen M. Assessing surgeon-level risk of patient harm during elective surgery for publica reporting. 2015. Available at: https://static.propublica.org/projects/patient-safety/methodology/surgeon-level-risk-methodology.pdf. Accessed December 17, 2015.
3. Friedberg MW, Provonost PJ, Shahian DM, et al. *A Methodological Critique of the ProPublica Surgeon Scorecard.* Santa Monica, CA: RAND Corporation; 2015.
4. Krell RW, Hozain A, Kao LS, et al. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg*. 2014;149:467–474.

5. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *J Am Coll Surg*. 2013;217:336–346.
6. Shiloach M, Frencher SK Jr, Steeger JE, et al. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *J Am Coll Surg*. 2010;210:6–16.
7. *ACS NSQIP Semiannual Report.* Chicago, IL: American College of Surgeons; 2015.
8. Merkow RP, Ju MH, Chung JW, et al. Underlying reasons associated with hospital readmission following surgery in the United States. *JAMA*. 2015;313:483–495.
9. Pierce O, Allen M. ProPublica Surgeon Scarecard Appendices. 2015. Available at: https://static.propublica.org/projects/patient-safety/methodology/surgeon-level-risk-appendices.pdf. Accessed December 17, 2015.
10. Hall BL, Huffman KM, Hamilton BH, et al. Profiling individual surgeon performance using information from a high-quality clinical registry: opportunities and limitations. *J Am Coll Surg*. 2015;221:901–913.
11. Jaffe TA, Hasday SJ, Dimick JB. Power outage-inadequate surgeon performance measures leave patients in the dark. *JAMA Surg*. 2016. Epub ahead of print.
12. Singh S, Lin YL, Kuo YF, et al. Variation in the risk of readmission among hospitals: the relative contribution of patient, hospital and inpatient provider characteristics. *J Gen Intern Med*. 2014;29:572–578.
13. Sutherland T, David-Kasdan JA, Beloff J, et al. Patient and provider-identified factors contributing to surgical readmission after colorectal surgery. *J Invest Surg*. 2016;1–7.
14. Dahlke AR, Chung JW, Holl JL, et al. Evaluation of initial participation in public reporting of American College of Surgeons NSQIP surgical outcomes on Medicare's Hospital Compare website. *J Am Coll Surg*. 2014;218:374–380.
15. Adams J. *The Reliability of Provider Profiling.* Santa Monica, CA: RAND Corporation; 2009.
16. National Quality Forum Measure Testing Task Force. Guidance for Measuring Testing and Evaluating Scientific Acceptability of Measure Properties. 2011. Available at: http://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Measure_Testing_Task_Force_Final_Report.aspx. Accessed April 4, 2016.

DISCUSSANTS

P. Angelos (Chicago, IL):

Thank you for the opportunity to discuss this paper. I congratulate the authors on this important study that was really very nicely presented.

Dr Ban and colleagues, Dr Bilimoria, have carefully analyzed the ProPublica surgeons scorecard of individual surgeon outcomes and compared it with ACS NSQIP data to assess the value of the score card data. The authors have shown that ProPublica is providing not only an inaccurate picture, but also an incomplete picture of the data, and the study, I think, clearly shows the challenges in setting up any type of individual surgeon scorecard and the impact of the choices about what procedures to include and whether inpatient or outpatient procedures are included will have on the results.

So, I have a few questions that are perhaps not surprisingly more philosophical in nature.

As surgeons, I think we are all interested in obtaining good informed consent from our patients so that we can improve the quality of the decisions that our patients make. So do you believe that accurate individual surgeon outcomes are important data for patients to have access to when making decisions about whether to have surgery from a particular surgeon?

Second, if you do not believe that individual surgeon outcomes should be tracked and reported to patients, why do you think this information is not important for patient decision making? And if you do think that individual surgeon outcomes are important for patients to know, now that you have shown that ProPublica is doing such a bad job, why is NSQIP not stepping in and providing this data?

Finally, I think ProPublica has tried, although ineffectively, to meet a perceived demand on the part of the public to know this information. So I think the big question, I believe, is do we as surgeons have a responsibility to collect and communicate accurate individual surgeon outcome data?

Response From K.Y. Bilimoria:

First, I think there is a clear need and a demand from the public that we provide individual surgeon-level performance data. ProPublica created this new outcome measure which was previously undescribed. If they had simply used readmissions or mortality, well-tested measures, and had used proper methods, we would be having a much different debate. We would be having a nice debate. It would be a debate about whether the attribution should occur at the individual surgeon level or at the hospital level, and that is a good debate. I would like to have had that debate. Instead we are left discussing methodological flaws. But we certainly have a responsibility to provide information to the public about surgeon performance.

So where should we go from here? We could go with process measures. Those are easier to assess at the individual surgeon level, smaller numbers, no risk adjustment is typically needed, and easily actionable. So process measures would be one good option.

The other important option is patient experience measures. We can also do some outcomes measures, but they need to be well tested and validated. We need to stick with things that we have been working on, testing, and iteratively improving. And if we are going to do something new, it should be well tested prior to public release.

So I absolutely think we should continue to provide more data to patients at the individual surgeon level. There is a demand for it, and we should lead the charge.

As for NSQIP, NSQIP is intended to be a hospital-based quality registry. The sampling is not set up to capture a large number of cases, necessarily, per surgeon. In some cases it can be done. Bruce Hall published a nice paper a few months ago looking at what is required to get reliable performance estimates. The issue of reliability and small sample sizes is a huge issue. So I think that for certain outcomes and for certain cases, we can report outcomes. But, again, the real focus of NSQIP is for internal quality improvement, and it does that really well, and as soon as we make it a public reporting issue, it may change how the data are perceived and reported. And so I think that should be done very carefully and cautiously as we move forward.

K.C. Kent (Madison, WI):

I just want to echo some of Peter's comments. I think this is coming. There is no doubt that ProPublica has not done it well. There is little doubt that NSQIP can do it much better.

I do think we as surgeons will need to take the lead and come up with something that works. The problem is that although NSQIP is a great tool, most of the surgeons in this country do not have access to NSQIP.

The question is, if you just have Medicare data available, or state databases, what are the factors that could be measured that you would project might be useful in differentiating a well-functioning surgeon from a surgeon that does not perform so well?

Response From K.Y. Bilimoria:

I think we are very limited with what we can do with the administrative data. I think we can look at mortality, readmissions—some of these hard endpoints. However, there may not be a reliable differences in performance between surgeons because mortality rates and case volumes are pretty small generally for individual surgeons.

Moving beyond that, I think it is on us to be able to collect better measures that are measurable across the entire country, and

some of the new CMS MACRA requirements for individual surgeon reporting will help address those issues. We may have some more robust performance measures, process measures for individual surgeons, that we can then use to provide some of the information to patients.

Certainly, a number of systems have gone toward publicly reporting individual surgeon/patient experience scores, and I think the hallmark for that is University of Utah. They have had tremendous improvement in performance by doing that, and it does provide useful information for patients.

D. Fry (Chicago, IL):

I must declare my conflicts of interest here since I am a consultant to Consumer Reports, to the Empirica Corporation, and to the Center for Special Services in evaluating provider performance.

Having said that, I would like to sort of take the middle road of suggesting that perhaps NSQIP and ProPublica are missing the mark by virtue of how complications are being evaluated.

ProPublica basically turned their back on it, and that is because it is hard to believe when you look at coded records. NSQIP codes who checks the box and probably is overreporting serious complications of care.

So what our own studies were, we used prolonged length of stay as a risk-adjusted measure using control charts, we find that about 30% to 50% of coded complications—and I suspect boxes checked at NSQIP—in fact are discharged on time and the patients are never readmitted and they have actually a fairly uneventful recovery, meaning that early intervention is very significant.

Relative to mortality, we find that there are 3 times as many deaths in the 90 days following discharge as occurred in the hospital, and that mortality rates are now underreported, probably, by NSQIP, and certainly by ProPublica.

So I guess, Karl, I would like to ask you, what is the best way for us to measure a complication of care?

I think the reason this entire subject is sort of wallowing between extremes is that I do not think we have good measures of outcomes. So what is going to be the best way of doing this, particularly to capture then the fact that more complications are declared after the patient is discharged than actually occurred during the hospitalization?

Response From K.Y. Bilimoria:

I think I would have to take issue with the fact that we do not have good outcome data, and I would have to disagree that NSQIP uniformly either undercalls or overcalls complications. I think validations and audits of ACS NSQIP data have shown that the data are accurate. We do a lot of work to ensure the standardization of collection of all data, particularly the 30-day outcomes. And I think that NSQIP and the STS registry are the gold standard for surgical registry data.

I think you bring up an interesting point to consider. One of the consequences of this measure, the way that ProPublica has done it, is that they have not captured the inpatient complications and that is where most of the complications occur. If a surgeon has a lot of early inpatient complications, that surgeon could theoretically look better than a surgeon with fewer inpatient complications on ProPublica's flawed outcome measure. Moreover, if one surgeon has a shorter length of stay, more complications will actually occur during the outpatient setting compared to a surgeon who has a longer length of stay and has inpatient complications. So this may be an example of a paradoxical measure where better-performing surgeons may actually appear to be poor performers just because of the way the measurement is done.

J.F. Burdick (Baltimore, MD):

I should disclose that I just published a book on health care reform. And this is not an advertisement; it's just legally necessary.

But Medicare billing data are a standard way to look at things. And the big exception to that really is NSQIP, and I think a shoutout to us as surgeons for the excellent system we have got, and it is recognized by many people. But you need clinically relevant data rather than billing data to really look at this.

So the vision to propose is that we have a national electronic medical records system based entirely on clinical information so it's easy and useful. Not billing. Billing can be dealt with by an app on the electronic medical records billing system nationally.

Some of the things that are complained about in the other systems might help, so the questions is, what do you think, and have you thought in doing this? If only you just had all the billing information, and not all the clinically relevant information, lab tests and so forth, and not the billing data to work with, do you think it would make a big difference?

Response From K.Y. Bilimoria:

I think that is the holy grail: to be able to automatically extract data from the EHR and get it uniformly from across the country. It is challenging to do for numerous reasons, but we are certainly working on it.