RAND CORPORATION

Perspective

Expert insights on a timely policy issue

A Methodological Critique of the ProPublica Surgeon Scorecard

Mark W. Friedberg, Peter J. Pronovost, David M. Shahian, Dana Gelb Safran, Karl Y. Bilimoria, Marc N. Elliott, Cheryl L. Damberg, Justin B. Dimick, Alan M. Zaslavsky

Summary

On July 14, 2015, ProPublica published its *Surgeon Scorecard*,[1] an online tool that displays "Adjusted Complication Rates" for individual, named surgeons for eight surgical procedures performed in hospitals.

Public reports of provider performance (or, *performance reports*) have the potential to improve the quality of health care that patients receive. Valid performance reports (i.e., reports that truly measure what they are advertised as measuring) can stimulate providers to make quality improvements and can help patients make better selections when choosing among health care providers. However, performance reports with poor measurement validity and reliability are potentially damaging to all involved. Therefore, it is important to critically examine the methods used to produce any performance report.

Measuring provider performance is challenging, but methods exist that can help ensure that performance reports are valid and display true differences in performance. This methodological critique of the ProPublica *Surgeon Scorecard* has three goals: to explain methodological issues in the *Scorecard*, to suggest ways in which the *Scorecard* can be improved, and to inform the public about these aspects of the *Scorecard*. An overview of our conclusions with respect to the first two goals follows. The third—to inform the public—exists because the *Scorecard* is currently available to the public, and, based on our critique, we hope patients who are choosing a surgeon will be better able to decide how much weight to give the data presented in the *Scorecard*.

Summary (continued)

Methodological Issues in the *Scorecard*:

- **The "Adjusted Complication Rates" reported in the *Scorecard* are not actually complication rates.** Instead, the "Adjusted Complication Rate" is a combination of hospital readmissions for conditions plausibly related to surgery (93 percent of events) and deaths (approximately 7 percent of events) within 30 days. However, most serious complications occur during the index admission, and many complications occur within 30 days post-discharge but without a readmission, or occur beyond the 30-day period. Other than death, none of these complications—many of which represent the most significant surgical risks and greatest detriment to patient long-term quality of life (such as urinary incontinence or erectile dysfunction following radical prostatectomy)—is included in the *Scorecard*. Most importantly, failure to include complications occurring during the index hospitalization is an unprecedented and untested departure from usual practices in measuring surgical complications, and one that undoubtedly results in a large proportion of serious surgical complications being excluded from the ProPublica measure.
- **As currently constructed, the *Scorecard* masks hospital-to-hospital performance differences, thereby invalidating comparisons between surgeons in different hospitals.** By setting the hospital random effects equal to 0 in calculating the "Adjusted Complication Rates," the ProPublica *Surgeon Scorecard* masks hospital-to-hospital variation that actually is present (according to ProPublica's models), thereby misleading patients in a systematic, albeit unintended, fashion. Put another way, the current *Scorecard* methodology ignores any hospital-level performance variation that would reflect (a) critical aspects of care that are intrinsic to a hospital, such as the adequacy of anesthesia staff, nursing, infection control procedures, or equipment, or (b) systematic recruitment of surgeons with superior (or inferior) skills by hospitals.
- **The accuracy of the assignment of performance data to the correct surgeon in the ProPublica *Surgeon Scorecard* is questionable.** Claims data, which form the basis of the *Scorecard*, are notoriously inaccurate in individual provider assignments, and the *Scorecard*, as originally published, included case assignments to nonsurgeons and to surgeons of the wrong subspecialty. There is reason to suspect that these readily detectable misattributions are symptoms of more pervasive misattributions of surgeries to individual surgeons, and that these errors are still present in the *Scorecard*.
- **The adequacy of the *Scorecard's* case-mix adjustment is questionable.** The aggregate patient "Health Score," which reflects ProPublica's overall estimate of inherent patient risk (and the only such estimate for three of the eight reported surgical procedures), has a coefficient estimate of 0, meaning this patient risk score has no effect in ProPublica's risk-adjustment models. A likely explanation is that ProPublica's case-mix adjustment method fails to capture important patient risk factors. None of ProPublica's methods accounts for the risk factors present in more-detailed surgical risk models derived from clinical data.

Summary (continued)

- **The *Scorecard* appears to have poor measurement reliability (i.e., it randomly misclassifies the performance of many surgeons).** Measurement reliability, which assesses the ability to distinguish true differences in performance between providers, is a key determinant of provider performance misrepresentation due to chance and should be calculated for all performance reports. Calculating reliability is particularly critical when measuring the performance of individual providers, where the number of cases used to rate a provider's performance can be quite small. Based on the width of the confidence intervals presented in the *Scorecard*, measurement reliability appears to be quite low for the vast majority of surgeons, with random misclassification rates between the *Scorecard's* implied risk classes (low, medium, and high "Adjusted Complication Rates") approaching 50 percent for some surgeons.

Ways to Improve the *Scorecard*:

- **Rename the "Adjusted Complication Rate" measures reported in the *Scorecard*.** Using a name that is more indicative of what is actually being measured will reduce the risk that *Scorecard* users (e.g., patients and providers) will misinterpret data on display, for example, by believing that a surgeon with a relatively low "Adjusted Complication Rate" has a lower overall rate of complications (in-hospital, post-discharge, and long-term) for a given procedure than a surgeon with a higher "Adjusted Complication Rate." In addition, ProPublica could attempt to perform scientifically credible validation of "Adjusted Complication Rates" as measures of true complication rates. If such efforts fail to validate the current measures, they might identify ways to improve these measures substantially.
- **Correct the statistical method for handling hospital contributions to the individual surgeon performance data presented in the *Scorecard*.** Setting hospital random effects equal to 0 is a methodological decision with no good justification, and it should be corrected in the existing *Scorecard*.
- **Validate the assignment to individual surgeons of the surgeries, readmissions, and deaths that are counted in the "Adjusted Complication Rates."** A validation study, comparing claims-based surgeon assignments with those derived from medical records for a representative sample of surgeries, could determine the extent of misattributed events. An informed judgment about whether the rate of misattribution is acceptable for high-stakes public reporting could then be made.
- **Validate the case-mix adjustment methods used to generate "Adjusted Complication Rates" for each surgeon.** Questions about the adequacy of risk adjustment could be addressed by methodologically rigorous validation, preferably using robust clinical registry data that exist for several of the ProPublica procedures. This exercise might also lead to the conclusion that stronger case-mix adjustment methods are needed to enable fair comparisons between providers.

Summary (continued)

- **Specify minimum acceptable thresholds for measurement reliability and abide by them.** State-of-the-art performance reports require minimum reliability to be achieved before publicly reporting performance data, and such reports warn their users when reliability is low. Having no minimum measurement reliability criterion for performance reporting is a departure from best practice, and one that appears to impose a high risk of both misclassifying individual surgeons and misdirecting report users.
- **Eliminate the implicit categorization of surgeons as having low, medium, or high "Adjusted Complication Rates."** The distinctions between these categories lack inherent meaning and, as described above, appear to have exceedingly high random misclassification rates for many surgeons. As a consequence, the red exclamation points marking hospitals with one or more surgeons having high "Adjusted Complication Rates" also should be eliminated.

Conclusion

ProPublica's stated goals in producing the *Surgeon Scorecard* are laudable: "to provide patients, and the health care community, with reliable and actionable data points, at both the level of the surgeon and the hospital, in the form of a publicly available online searchable database."[2] However, as with any performance report, the *Scorecard's* ability to achieve these goals is limited by the rigor of the methods and the adequacy of the underlying data. Our critique of the ProPublica *Surgeon Scorecard* has identified substantial opportunities for improvement. Until these opportunities are addressed, we would advise users of the *Scorecard*—most notably, patients who might be choosing their surgeons—not to consider the *Scorecard* a valid or reliable predictor of the health outcomes any individual surgeon is likely to provide.

It is important to advise patients to ask all prospective surgeons about the risks of poor surgical outcomes, for hospitals to monitor the quality of their staff members (including surgeons), and for providers to substantially improve their efforts to collect and share useful performance data publicly. We hope that publication of the ProPublica *Surgeon Scorecard* will contribute to these broader efforts, even though there is substantial reason to doubt the *Scorecard's* current usefulness as a source of information about individual surgeons' quality of care. We also hope that this critique of the methods underlying the ProPublica *Surgeon Scorecard* will contribute to the development of more-valid and reliable performance reports in the future.

Introduction

Public reports of provider performance (or, *performance reports*) have the potential to improve the quality of health care that patients receive. A valid performance report can spur provider quality improvement and usefully inform patients' choices of providers.

Reports on provider performance are estimates, drawn from finite samples of past performance, of providers' underlying "true" performance as it might be experienced by future patients.[3, 4] If these reports are accurate, patients can use them to choose better providers (i.e., those providers who are most likely to improve their health outcomes) and avoid worse performers. By seeing their own results benchmarked against those of their peers caring for a similar mix of patients, providers can learn which of the services they offer need improvement. Properly conducted, public reporting fulfills the ethical obligations of facilitating patient autonomy in decisionmaking and promoting better patient outcomes (or, *beneficence*).[5–8]

Conversely, methodologically flawed, inconsistent, or invalid report cards might cause substantial harm by misleading or confusing patients and incorrectly classifying providers.[9–16] Such reports could lead patients to choose or avoid providers based on inaccurate information. Patients are thereby harmed in the short term by making worse choices of providers and in the long term by misinformed or inappropriately incentivized providers, who might fail to make necessary improvements or might avoid patients who are appropriate for a surgical procedure but who pose a greater risk of generating a poor reported outcome.

Furthermore, poorly designed performance reports can cause unjustified damage to provider reputations and foster cynicism about or mistrust of other, more-valid reports, thereby undermining provider engagement in quality-improvement efforts and the entire quality-improvement enterprise.

No report can be perfectly valid and reliable. Case-mix adjustment cannot ensure completely fair comparisons, and some degree of measurement error is unavoidable. Some imperfections in and limitations of performance reports are generally accepted when they are publicly acknowledged and when their benefits (e.g., the number of patients directed to better providers) substantially outweigh their harms (e.g., the number of patients misdirected to worse providers). But a misleading report based on seriously flawed methodology can harm both patients and providers. Therefore, it is important to critically examine the methods used to produce a performance report.

This methodological critique of the ProPublica *Surgeon Scorecard* has three goals: to explain methodological issues in the *Scorecard*, to suggest ways in which the *Scorecard* can be improved, and to inform the public about these aspects of the *Scorecard*. The first two of these goals are standard for scientific peer review. Our third goal—to inform the public—exists because the *Scorecard* has been published already and is the reason we are publishing this critique openly rather than communicating it only to ProPublica, which we have already done. Based on this critique, we hope patients who are choosing a surgeon will be better able to decide how much weight to give the data presented in the *Scorecard*.

This critique is based on the following sources of information about the *Scorecard*:

- *Assessing surgeon-level risk of patient harm during elective surgery for public reporting: Whitepaper as of August 4, 2015,* by Olga Pierce and Marshall Allen of ProPublica (henceforth, "the white paper"), and its appendix[2, 17] [accessed September 9, 2015]

- The online display of surgeon-specific "Adjusted Complication Rates" by ProPublica[1] [accessed September 9, 2015]
- Correspondence received from ProPublica on August 3, August 7, August 13, August 15, and September 11, 2015, in response to our questions about methods underlying the *Scorecard*.

Brief Description of the ProPublica *Surgeon Scorecard*

The online *Surgeon Scorecard* tool published by ProPublica on July 14, 2015, reports the performance of individual, named surgeons on a new measure called "Adjusted Complication Rate" for each of eight surgical procedures: laparoscopic cholecystectomy; radical prostatectomy; transurethral prostatectomy; cervical fusion of the anterior column, anterior technique; lumbar and lumbosacral fusion of the posterior column, posterior technique; lumbar and lumbosacral fusion of the anterior column, posterior technique; total hip replacement; and total knee replacement.[1] The data underlying the report are Medicare 100% Standard Analytic Files for 2009–2013, which reflect hospital care for traditional (fee-for-service) Medicare beneficiaries.

For each of the eight surgical procedures, the corresponding "Adjusted Complication Rate" is a composite measure of death within 30 days of surgery (approximately 7 percent of events[i]) and hospital readmissions occurring within 30 days of the initial hospital discharge, for which the primary admitting diagnosis was plausibly related to the index surgery (approximately 93 percent of events counted by the ProPublica *Surgeon Scorecard*). Deaths and readmissions are weighted equally. Other than deaths, the "Adjusted Complication Rate" does not include any complications occurring during the index admission, complications occurring after discharge but not accompanied by hospital readmission, or any complications occurring after 30 days have elapsed. This new measure, created by ProPublica, has not been used in any other performance report.

The ProPublica *Surgeon Scorecard* displays point estimates and confidence intervals for each surgeon on the "Adjusted Complication Rate" measure for the corresponding procedure, which ProPublica has calculated using hierarchical linear models described in *Assessing surgeon-level risk of patient harm during elective surgery for public reporting: Whitepaper as August 4, 2015*, by Olga Pierce and Marshall Allen of ProPublica.[2] These hierarchical linear models have a logistic-normal functional form. The independent variable (outcome) is a binary indicator for events counted in the "Adjusted Complication Rate" measure. The independent variables (predictors) are random effects for hospitals and surgeons and fixed effects for patient-level risk adjustment: age, sex, "Health Score" (based on the VanWalraven modification of the Elixhauser index), and one additional variable for five of the eight reported surgical procedures.[ii]

Using these estimates, the *Surgeon Scorecard* classifies the "Adjusted Complication Rate" of each surgeon who performed 20 or more of the indicated procedure as low, medium, or high using thresholds specific to each procedure. The *Scorecard* also applies

[i] Including the 910 patients who died after a qualifying readmission.

[ii] The additional risk adjustment variables are performance of multilevel spinal fusion (for the three spinal fusion surgeries), pancreatitis diagnosis (for cholecystectomy), and an indicator for whether surgery was robot-assisted (for radical prostatectomy).

exclamation points beside the names of hospitals containing at least one surgeon with a high "Adjusted Complication Rate."

In generating each surgeon's "Adjusted Complication Rate" for a given procedure, ProPublica has set the hospital random effects equal to 0 (i.e., "presuming that the surgeon is operating at an average hospital," as explained in the white paper),[2] regardless of the hospital random effect values actually estimated by the corresponding hierarchical linear model.

Methodological Issues in the ProPublica Surgeon Scorecard, and Suggestions for Improvement

The "Adjusted Complication Rates" Reported in the Scorecard Are Not Actually Complication Rates

Instead, the "Adjusted Complication Rate" is an *adjusted 30-day readmission rate* for conditions plausibly related to surgery (approximately 93 percent of events), *plus a relatively small number of deaths* (approximately 7 percent of events). The "Adjusted Complication Rate" does not include complications in multiple categories that matter to patients: complications occurring during the index hospitalization, even those that are quite serious (except death); complications occurring after discharge but not accompanied by hospital readmission; or any complications occurring after 30 days have elapsed post-discharge.

Failure to include complications occurring during the index hospitalization (i.e., the hospital admission in which the surgery took place) is an unprecedented and untested departure from usual practice, and one that results in a large proportion of surgical complications being excluded from the ProPublica measure. Specifically, detailed clinical data from the National Surgical Quality Improvement Program (NSQIP) have demonstrated that approximately two-thirds (67.1 percent) of 30-day surgical complications occur during the index admission.[18]

In addition, by not including long-term complications, the "Adjusted Complication Rate" does not reflect some of the most common complications (and most serious complications, excepting death) of the measured surgeries. For example, following radical prostatectomy, rates of erectile dysfunction (more than 50 percent by some estimates) and urinary incontinence (approximately 14 percent with residual leakage),[19] even at two years postsurgery, are far higher than the mean "Adjusted Complication Rate" for this procedure (2.9 percent), and such complications will have far greater lifelong consequence to the patient than a readmission for a fever (one of the qualifying conditions used in the ProPublica measure). Whether the readmissions captured in ProPublica's "Adjusted Complication Rates" have any empirical association with long-term health outcomes is unknown.

Recent analyses of NSQIP data have found statistically significant associations between the occurrence of post-discharge complications and readmission rates within 30 days of surgery.[20] However, such associations do not constitute evidence that surgeon-level readmission rates are valid measures of surgeon-level complication rates. To know the validity of readmission rates as measures of complication rates, we would need to know their sensitivity (the percentage of post-discharge complications that result in readmission) and specificity (the percentage of readmissions that are due to complications). While only including readmissions for diagnoses consistent with surgical complications is likely to increase the specificity of the "Adjusted Complication Rate," ProPublica's method-

ology does nothing to increase the sensitivity of readmissions as indicators of complications.

There are no data on the sensitivities or specificities of the readmissions counted in ProPublica's "Adjusted Complication Rates" as indicators of 30-day post-discharge complications. Moreover, there is likely to be substantial surgeon-to-surgeon variation in the sensitivity of readmissions as measures of such complications; this is because of factors including but not limited to differences in patient case-mix, differences in resources allowing management of complications in the outpatient rather than inpatient setting, and differences in surgeons' innate propensities to readmit a given patient (for any given complication). Surgeon-to-surgeon differences on any of these characteristics would undermine the validity of individual surgeons' readmission rates as measures of their complication rates, relative to other surgeons.

Suggestion for Improvement

To address these methodological concerns, the "Adjusted Complication Rate" measure should be renamed. Giving the measure a more accurate name would reduce the risk that *Scorecard* users (e.g., patients and providers) would misinterpret data on display—preventing, for example, casual readers from believing that "Adjusted Complication Rates" actually reflect surgeons' overall rates of complications for the procedures in question. A more substantive change would be to perform scientifically credible validation of "Adjusted Complication Rates" as measures of complications. Sensitivity and specificity (and especially surgeon-to-surgeon variation in sensitivity) could be calculated by comparison to a gold-standard clinical dataset for a representative sample of surgeons. This validation exercise might identify ways that the current measure could be improved.

As Currently Constructed, the *Scorecard* Masks Hospital-to-Hospital Performance Differences, Thereby Invalidating Comparisons Between Surgeons Who Operate in Different Hospitals

For each measure, the ProPublica *Surgeon Scorecard* reports an "Adjusted Complication Rate" for each surgeon, defined in the white paper as "presuming that the surgeon is operating at an average hospital (this is achieved by setting the hospital random effect to 0)." Later in the white paper, this methodological choice is restated: "The Adjusted Complication Rate does not directly represent a surgeon's past outcomes. It is an assessment of how s/he would perform at a hypothetical average hospital, on a standardized patient pool."[2]

This is a highly problematic methodological choice if the intent of the report is to help patients choose from among surgeons located in different hospitals. Patients cannot receive care from a chosen surgeon in a "hypothetical average hospital" but must receive care in one of the hospitals (often few or even a single hospital) where the selected surgeon actually provides the desired service. ProPublica's models reveal substantial between-hospital performance variation that is not explained by individual surgeon effects (see Tables 4 and 8 in the white paper).[2] By setting the hospital random effects equal to 0, the ProPublica *Surgeon Scorecard* is masking such hospital-to-hospital variation. This is especially important because nonzero hospital random effects might capture critical aspects of care that are intrinsic to the hospital and can affect outcomes, such as the adequacy of anesthesia staff, nursing,

infection control procedures, or equipment. Thus, by excising one part of the inseparably connected hospital and surgeon contributions to quality, the ProPublica *Surgeon Scorecard* underestimates the performance that patients are likely to receive in truly high-performing hospitals and overestimates performance in truly low-performing hospitals.

In addition, hospital random effects can in part represent hospitals' systematic recruitment of surgeons with superior skills (or conversely, systematic recruitment of inferior surgeons). When such surgeons practice exclusively (or primarily) in a single hospital, the estimates reported in the *Scorecard* would then represent performance relative to a superior (or inferior) within-hospital comparison group, rather than to a uniform national standard. Thus, the "hypothetical average hospital" created by setting hospital random effects equal to 0 might in many cases be "average" primarily in the sense that hospital-level performance differences caused by especially good (or poor) recruiting of individual surgeons are attenuated. ProPublica's approach thus might underestimate (or overestimate) performance for all surgeons within a hospital, even when their collective superior (or inferior) performance solely reflects the average of their intrinsic individual performance. Therefore, it is unlikely that setting hospital random effects equal to 0 will lead to valid identification of surgeons who are national outliers as individuals (which is the other conceivable justification for this methodological choice).

As a corollary, the problems created by ProPublica's decision to set hospital random effects equal to 0 are magnified when red exclamation points representing surgeons with low-ranking random effects are applied to hospital displays. This creates a de facto hospital measure from an analysis that explicitly omits an important component of hospital-to-hospital performance variation.

As a second corollary, ProPublica has probably understated the importance of choosing the right hospital given the inextricable linkage between hospital and surgeon performance. The ProPublica article "Making the Cut: Why Choosing the Right Surgeon Matters Even More Than You Know" states that "it is much more important to pick the right surgeon" than to pick the right hospital.[iii, 21] This conclusion stems from results summarized in the white paper abstract: "Finally, a comparison of the standard deviations of hospital and surgeon random effects found that surgeon performance accounts for more of the variability of performance between hospitals than hospital-wide performance on a given procedure."[2] This statement implicitly interprets hospital random effects as representing nonsurgeon aspects of quality and as being entirely distinguishable from surgeon random effects representing individual surgeon performance. However, for the reasons detailed above, hospital and surgeon contributions to performance ("effects" in the common use of the term) are not actually distinguishable in

[iii] The full quote is:

> It's conventional wisdom that there are "good" and "bad" hospitals—and that selecting a good one can protect patients from the kinds of medical errors that injure or kill hundreds of thousands of Americans each year. But a ProPublica analysis of Medicare data found that, when it comes to elective operations, it is much more important to pick the right surgeon.

Moreover, the White Paper includes the following statements:

> . . . overall, hospitals (as isolated from the surgeons who perform surgeries there) are relatively similar. When hospitals have different adjusted complication rates, it is mainly due to variation in surgeon performance. Likewise, a patient's choice of hospital will, in general, have less impact on his or her risk of readmission or death than choice of surgeon.

ProPublica's models. Therefore, an interpretation that choosing a surgeon is more (or less) important than choosing a hospital is not supported by ProPublica's analytic approach.

Suggestion for Improvement

To address methodological concerns about the handling of hospital contributions to performance, ProPublica should include the hospital random effects estimated by its hierarchical linear models in each surgeon's predicted "Adjusted Complication Rate." Setting hospital random effects equal to 0 in making these predictions is a methodological decision with no good justification, and it should be corrected in the existing *Scorecard*.

The Accuracy of the Assignment of Performance Data to the Correct Surgeon in the ProPublica *Surgeon Scorecard* Is Questionable

The ProPublica *Surgeon Scorecard* reports the performance of individual, named surgeons. Therefore, accurate assignment of clinical events contributing to the denominator and numerator of the "Adjusted Complication Rate" is critical to the validity of the *Scorecard*.

Our concerns about assignment were spurred by finding, upon review of convenience samples (i.e., individuals looking at their own hospitals' *Scorecard* data), that the ProPublica *Surgeon Scorecard* published on July 14, 2015, included physicians who are not surgeons or who never perform the surgeries that the *Scorecard* claims they have performed. For example, at Massachusetts General Hospital,[iv] the ProPublica *Surgeon Scorecard* listed four nonsurgeons (a cardiologist, a pulmonary and critical care specialist, and two general internists) as surgeons who have performed total knee replacements fewer than 20 times and another two inapplicable physicians (an interventional cardiologist and a heart surgeon) as having performed hip replacements. ProPublica has since identified and removed 66 nonsurgeons from the *Scorecard* using the provider taxonomy variables (i.e., physician specialties) available in the Medicare Standard Analytical Files; cases formerly assigned to nonsurgeons are now listed as having "unknown" surgeon in the *Scorecard*.[v]

While eliminating readily identifiable attribution errors (which stem from hospital claims submission errors) may remove some data problems, these obvious misattributions might be the tip of a still-uncorrected iceberg of pervasive errors in the source data. In other words, the inaccurate assignment of measured events to nonsurgeons is primarily concerning as a signal that more-extensive misattributions (between surgeons of the same subspecialty within the same hospital) might also be present in the claims data.

As others have noted,[22] a 2012 study sponsored by the Centers for Medicare & Medicaid Services (CMS) found that the underlying rate of operating National Provider Identifier (NPI) mismatch between Medicare part A and part B claims (which could indicate misattributed cases in hospital claims)[vi] for four examined surgical procedures exceeds 28 percent of surgeries.[23]

[iv] We thank Dr. Ishani Ganguli and Dr. David Shahian for pointing this out.

[v] Written communication with Stephen Engelberg and Olga Pierce, August 7, 2015. In addition, Stephen Engelberg reported on September 11, 2015, that ProPublica has decided to remove from the *Scorecard* all surgeons whose hospitals misattributed (to nonsurgeons or surgeons of inapplicable subspecialties) more than 100 claims overall or more than 5 percent of the claims in one procedure.

[vi] Use of multiple NPIs for the same surgeon (e.g., the surgeon's individual NPI and the NPI of the group to which the surgeon belongs) also can create NPI

Suggestion for Improvement

This methodological concern can be addressed by validating the assignment to individual surgeons of the surgeries, readmissions, and deaths that are counted in the "Adjusted Complication Rates." A validation study comparing claims-based assignments to medical records for a representative sample of surgeries could determine the extent of misattributed events and enable informed judgment about whether the rate of misattribution is acceptable for high-stakes public reporting. A logical first step toward validation would be to allow a period of internal review and confirmation by the providers, as CMS and other oversight agencies have done before releasing certain provider information to the public.

The Adequacy of Case-Mix Adjustment in the *Scorecard* Is Questionable

For outcome measurement, comparisons among providers are misleading if they do not account for differences in the health status and risk factors of the patients who receive care from these providers. Case-mix adjustment (or, *risk adjustment*) refers to any procedure intended to remove the effect of such differences in patient characteristics and thereby enable fair comparisons among providers. To perform case-mix adjustment for the *Scorecard*, ProPublica's hierarchical linear model includes fixed effects for patient-level risk adjustment: age, sex, "Health Score" (based on the Elixhauser index), and one additional variable for five of the eight reported surgical procedures.[vii]

As shown in Tables 4 and 5 of the white paper, the aggregate patient "Health Score" has a coefficient estimate of 0—i.e., it has no effect in ProPublica's risk-adjustment model.[2] Given major inter-provider differences in risk factor prevalence that have been demonstrated using studies based on audited clinical data even with highly homogeneous procedure categories,[24, 25] the most likely explanation for this lack of effect is that ProPublica's "Health Score" measure fails to capture important patient risk. In addition, none of ProPublica's risk adjustment algorithms contains the risk factors present in more-detailed surgical risk models derived from clinical data (e.g., indication for the operation, presence of ascites, body mass index, and American Society of Anesthesiologists class—all of which have been found to substantially affect adjusted performance in the NSQIP).[26–31]

Suggestion for Improvement

Questions about the adequacy of case-mix adjustment could be addressed by validation of these adjustment methods. Ideally, this could be accomplished through comparisons with results obtained using risk models derived from clinical registry data, which are available for several of the procedures included in the *Scorecard*. It is also possible to derive credible risk-adjustment models from administrative data, such as those used in the *Hospital Compare* models of 30-day mortality rates for acute myocardial infarction,

mismatches between part A and part B Medicare claims. The proportion of such mismatches that represents surgeon misattribution in hospital claims is unknown.

[vii] The additional risk adjustment variables are performance of multilevel spinal fusion (for the three spinal fusion surgeries), pancreatitis diagnosis (for cholecystectomy), and an indicator for whether surgery was robot-assisted (for radical prostatectomy). Whether to use a robot during radical prostatectomy, when one is available, can be influenced by both patient characteristics and surgeon choice of technique. Because robot use is not completely exogenous to surgeon decision-making and might lie on the causal pathway between surgeons' choices and occurrence of complications, the decision to include robot-assistance in case-mix adjustment is not straightforward.

heart failure, and pneumonia.[32–34] The developers of these measures validated their results against gold-standard data derived from clinical sources.[35] Following these best practices for public reporting would improve the methodological credibility of ProPublica's Surgeon Scorecard.

The *Scorecard* Appears to Have Poor Measurement Reliability (i.e., Randomly Misclassifies the Performance of Many Surgeons)

Measurement reliability (see box on following page for definition) is a key determinant of provider performance misrepresentation due to random measurement error (i.e., chance). The reliability of a performance measure *and* the display format (in particular, the handling of performance thresholds and the treatment of scores lying most proximal to these thresholds) affect the likelihood of misclassifying a provider's true performance. ProPublica does not appear to have followed best practices in attending to these core factors that drive the rate of performance misclassification due to chance.

For each surgical procedure, the ProPublica *Surgeon Scorecard* classifies each surgeon as having a low, medium, or high "Adjusted Complication Rate." Though ProPublica does not report the degree of misclassification, we believe, given the distribution of scores and confidence intervals in the appendix to the white paper,[2] that measurement reliability is likely to be low for the vast majority of surgeons included in the *Scorecard*.

In addition, the risk of misclassification over the low-to-medium and medium-to-high "Adjusted Complication Rate" thresholds appears to approach 50 percent for some surgeons (little better than a coin flip). It is well understood that for performance point estimates near a threshold, the risk of random misclassification is high; this problem can be addressed by using buffered benchmarks (i.e., approaches that classify a surgeon as "low" or "high" only when the probability of random misclassification is below a preset maximum).[36] However, the *Scorecard* does not use buffered benchmarks or similar direct approaches to limiting the risk of random misclassification near a performance threshold.[37]

Suggestion for Improvement

We suggest specifying minimum acceptable thresholds for measurement reliability and maximum acceptable thresholds for random misclassification, and then abiding by them. State-of-the-art performance reports require minimum reliability to be achieved before reporting performance data, and such reports warn their users when reliability is low. Having no minimum measurement reliability criteria for performance reporting would be a departure from best practice[4, 36–38] and one that is likely to result in random misdirection of many report users. Procedures for calculating measurement reliabilities and random performance misclassification rates are available.[39, 40]

Performance Thresholds in the ProPublica *Surgeon Scorecard* Lack Inherent Meaning

For each procedure, ProPublica uses thresholds to divide surgeons' performance on the corresponding "Adjusted Complication Rate" measure into low, medium, and high ranges (i.e., green, yellow, and red areas). These thresholds also underlie a de facto classification system for hospitals through ProPublica's application of exclamation points beside the names of hospitals containing at least one surgeon with a high "Adjusted Complication Rate." The white

Definition of *Reliability* and Random Performance Misclassification

Reliability is a key metric of the suitability of a measure for profiling because it describes how well one can confidently distinguish the performance of one physician from that of another. Conceptually, it is the ratio of signal to noise. The signal, in this case, is the proportion of the variability in measured performance that can be explained by real differences in performance [of physicians].[39]

In reports of provider performance, the noise is due to the randomness of outcomes of the particular collection of patients whom a particular provider treated during the period used to construct the measure. It broadly reflects considerations such as sample size (low case volumes often lead to low reliability), the relative differences in outcomes among patients within a given provider, and the differences in outcomes between providers.

The reliability of the measures will affect the accuracy of any conclusions that are drawn about comparisons among providers. In particular, many performance reports, including ProPublica's *Surgeon Scorecard*, classify provider performance on continuous measures into discrete categories (e.g., low, medium, and high). Reliability is very important in determining the risk of random performance misclassification (i.e., purely due to chance). For any given classification system, higher reliability will reduce the risk of random misclassification.[4, 37] Reliability also affects the accuracy of other methods of comparison, such as ranking of surgeons or pairwise comparisons.

Random misclassification does not have a predictable direction; it is like flipping a coin to reward some providers and penalize others. Some amount of random misclassification is unavoidable in performance measurement because there is random variation in all statistical estimates, including estimates of provider performance.

Patients, the public, policymakers and measure developers have a wide range of opinions about the acceptable level of misclassification in a performance report. In one survey, for example, most patients thought that it would be reasonable for a report to misclassify 5–20 percent of providers.[42] Other stakeholders might have higher or lower tolerances for misclassification due to chance, and, for many, the entire concept is not well understood.

Statistical techniques, such as empirical Bayesian shrinkage estimates in hierarchical linear models (like those used by ProPublica), can reduce the rate of misclassification due to chance to some extent. But what if random misclassification rates remain unacceptably high after their application? Report producers can decide to simply state that an acceptably accurate performance report could not be produced. Or, producers can publish a report with high random misclassification rates, under the assumption that confidence intervals or other representations of misclassification risk will be understandable to the general public. The problem with this latter approach is that, as research by Judith Hibbard and others has shown, such efforts are likely to be futile, as most people assume that the data contained in a published report are reliable.[43–45] Reporting unreliable information and then expecting people to discount it is a major departure from best practices in public reporting.[46]

paper reports that "The green, yellow, and red areas on these bars are determined by the shape of the distribution of adjusted surgeon complication rates for each procedure."[2] Unlike the *Hospital Compare* classification of hospitals on 30-day unplanned-readmission and death rates, the thresholds used in the ProPublica *Surgeon Scorecard* do not reflect statistically significant differences from the mean.[41]

Also, ProPublica's application of exclamation points to hospitals containing one or more surgeons with high a "Adjusted Complication Rate" has undesirable measurement properties. For two hospitals with identical average performance, the one with more surgeons is more likely to receive an exclamation point because it simply has more chances to contain a performance outlier. Therefore, ProPublica's method for assigning exclamation points is biased against larger hospitals.

Suggestion for Improvement

A straightforward solution to this problem would be to eliminate implicit categorization of surgeons as having low, medium, or high "Adjusted Complication Rates" and to remove the exclamation points altogether. This solution is especially attractive in light of the reliability and misclassification problems discussed above.

Conclusion

Given all of the methodological problems detailed herein, there is substantial reason to doubt that the ProPublica *Surgeon Scorecard*, as currently constructed, will help patients choose those surgeons who are most likely to provide good surgical outcomes.

ProPublica's goal in creating its *Surgeon Scorecard* is laudable and important: to improve the quality of surgical care that patients receive. The methodological challenges we have described in this critique are not unique to ProPublica's efforts. Designers of nearly all performance reports face them. But in light of these challenges and the serious potential unintended consequences of creating a misleading performance report, it is important for designers to carefully validate any new reporting methodology prior to publication, after which unintended damage to patients and providers is much more difficult to reverse. Recognizing the need to provide patients with more information about the quality of care, a performance report with poor or untested validity and reliability might misinform more than inform and hurt more than help.

References

1. Wei S, Pierce O, Allen M. Surgeon Scorecard. Online tool. ProPublica, 2015. As of September 11, 2015: https://projects.ProPublica.org/surgeons/

2. Pierce O, Allen M. Assessing surgeon-level risk of patient harm during elective surgery for public reporting (as of August 4, 2015). White paper. ProPublica, 2015. As of September 11, 2015: https://static.propublica.org/projects/patient-safety/methodology/surgeon-level-risk-methodology.pdf

3. Elliott M, Zaslavsky A, Cleary P. Are finite population corrections appropriate when profiling institutions? Health Services and Outcomes Research Methodology. 2006;6(3):153-6.

4. Friedberg MW, Damberg CL. A five-point checklist to help performance reports incentivize improvement and effectively guide patients. Health Aff. 2012;31(3):612-8.

5. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. New York: Oxford University Press; 2008.

6. Shahian DM, Edwards FH, Jacobs JP, Prager RL, Normand SL, Shewan CM, et al. Public reporting of cardiac surgery performance: Part 1. History, rationale, consequences. Ann Thorac Surg. 2011;92(3 Suppl):S2-11.

7. Shahian DM, Edwards FH, Jacobs JP, Prager RL, Normand SL, Shewan CM, et al. Public reporting of cardiac surgery performance: Part 2. Implementation. Ann Thorac Surg. 2011;92(3 Suppl):S2-11.

8. Clarke S, Oakley J. Informed Consent and Clinician Accountability: The Ethics of Report Cards on Surgeon Performance. Cambridge: Cambridge University Press; 2007.

9. Healthcare Association of New York State. HANYS' report on report cards: Understanding publicly reported hospital quality measures. Healthcare Association of New York State (HANYS), 2013 October.

10. Rothberg MB, Morsi E, Benjamin EM, Pekow PS, Lindenauer PK. Choosing the best hospital: The limitations of public quality reporting. Health Aff. 2008;27(6):1680-7.

11. Shahian DM, Wolf RE, Iezzoni LI, Kirle L, Normand SL. Variability in the measurement of hospital-wide mortality rates. N Engl J Med. 2010;363(26):2530-9.

12. Leonardi MJ, McGory ML, Ko CY. Publicly available hospital comparison web sites: Determination of useful, valid, and appropriate information for comparing surgical quality. Arch Surg. 2007;142(9):863-8; discussion 8-9.

13. Austin JM, Jha AK, Romano PS, Singer SJ, Vogus TJ, Wachter RM, et al. National hospital ratings systems share few common scores and may generate confusion instead of clarity. Health Aff. 2015;34(3):423-30.

14. Bilimoria KY, Chung J, Ju MH, Haut ER, Bentrem DJ, Ko CY, et al. Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure. JAMA. 2013;310(14):1482-9.

15. Rajaram R, Barnard C, Bilimoria KY. Concerns about using the patient safety indicator-90 composite in pay-for-performance programs. JAMA. 2015;313(9):897-8.

16. Rajaram R, Chung JW, Kinnier CV, Barnard C, Mohanty S, Pavey ES, et al. Hospital characteristics associated with penalties in the Centers for Medicare & Medicaid Services Hospital-Acquired Condition Reduction Program. JAMA. 2015;314(4):375-83.

17. Pierce O, Allen M. Assessing surgeon-level risk of patient harm during elective surgery for public reporting: Appendixes to white paper. ProPublica, 2015. As of September 11, 2015: https://static.propublica.org/projects/patient-safety/methodology/surgeon-level-risk-appendices.pdf

18. Bilimoria KY, Cohen ME, Ingraham AM, Bentrem DJ, Richards K, Hall BL, et al. Effect of postdischarge morbidity and mortality on comparisons of hospital surgical quality. Ann Surg. 2010;252(1):183-90.

19. Sanda MG, Dunn RL, Michalski J, Sandler HM, Northouse L, Hembroff L, et al. Quality of life and satisfaction with outcome among prostate-cancer survivors. N Engl J Med. 2008;358(12):1250-61.

20. Merkow RP, Ju MH, Chung JW, Hall BL, Cohen ME, Williams MV, et al. Underlying reasons associated with hospital readmission following surgery in the United States. JAMA. 2015;313(5):483-95.

21. Allen M, Pierce O. Making the cut: Why choosing the right surgeon matters even more than you know. ProPublica, 2015 July 13. As of September 11, 2015: https://www.propublica.org/article/surgery-risks-patient-safety-surgeon-matters

22. Dougherty G, Harder B. The U.S. News take on ProPublica's Surgeon Scorecard. US News and World Report. 2015 August 25.

23. Dowd B, Kane R, Parashuram S, Swenson T, Coulam RF. Alternative approaches to measuring physician resource use: Final report. 2012 April 9.

24. Shahian DM, He X, Jacobs JP, Rankin JS, Peterson ED, Welke KF, et al. Issues in quality measurement: Target population, risk adjustment, and ratings. Ann Thorac Surg. 2013;96(2):718-26.

25. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. Circulation. 2008;117(15):1955-63.

26. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: A decision aid and informed consent tool for patients and surgeons. Journal of the American College of Surgeons. 2013;217(5):833-42 e1-3.

27. Cohen ME, Ko CY, Bilimoria KY, Zhou L, Huffman K, Wang X, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: Patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. Journal of the American College of Surgeons. 2013;217(2):336-46 e1.

28. American College of Surgeons. National Surgical Quality Improvement Program: Semiannual Report. 2015 July 20.

29. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: Part 1. Coronary artery bypass grafting surgery. Ann Thorac Surg. 2009;88(1 Suppl):S2-22.

30. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: Part 2. Isolated valve surgery. Ann Thorac Surg. 2009;88(1 Suppl):S23-42.

31. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: Part 3. Valve plus coronary artery bypass grafting surgery. Ann Thorac Surg. 2009;88(1 Suppl):S43-62.

32. Bratzler DW, Normand SL, Wang Y, O'Donnell WJ, Metersky M, Han LF, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One. 2011;6(4):e17401.

33. Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation. 2006;113(13):1683-92.

34. Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation. 2006;113(13):1693-701.

35. Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, et al. Standards for statistical models used for public reporting of health outcomes: An American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation. Circulation. 2006;113(3):456-62.

36. Safran DG, Karp M, Coltin K, Chang H, Li A, Ogren J, et al. Measuring patients' experiences with individual primary care physicians: Results of a statewide demonstration project. J Gen Intern Med. 2006;21(1):13-21.

37. Friedberg MW, Damberg CL. Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting: A Guide for Community Quality Collaboratives. AHRQ Publication No. 11-0093. Rockville, MD: Agency for Healthcare Research and Quality. 2011 September.

38. Lyratzopoulos G, Elliott MN, Barbiere JM, Staetsky L, Paddison CA, Campbell J, et al. How can health care organizations be reliably compared? Lessons from a national survey of patient experience. Med Care. 2011;49(8):724-33.

39. Adams JL, Mehrotra A, McGlynn EA. Estimating Reliability and Misclassification in Physician Profiling. TR-863-MMS. Santa Monica, CA: RAND Corporation, 2010. As of September 11, 2015: http://www.rand.org/pubs/technical_reports/TR863.html

40. Adams JL. The Reliability of Provider Profiling: A Tutorial. TR-653-NCQA. Santa Monica, CA: RAND Corporation, 2009 As of September 11, 2015: http://www.rand.org/pubs/technical_reports/TR653.html

41. Centers for Medicare and Medicaid Services. 30-day unplanned readmission and death measures. Website. No date. As of July 30, 2015: https://www.medicare.gov/HospitalCompare/Data/30-day-measures.html

42. Davis MM, Hibbard JH, Milstein A. Issue Brief: Consumer Tolerance for Inaccuracy in Physician Performance Ratings: One Size Fits None. Issue Brief 110. Washington, D.C.: Center for Studying Health System Change, 2007 March.

43. Hibbard JH, Peters E, Slovic P, Finucane ML, Tusler M. Making health care quality reports easier to use. Jt Comm J Qual Improv. 2001;27(11):591-604.

44. Hibbard JH, Peters E. Supporting informed consumer health care decisions: Data presentation approaches that facilitate the use of information in choice. Annu Rev Public Health. 2003;24:413-33.

45. Hibbard JH, Greene J, Daniel D. What is quality anyway? Performance reports that clearly communicate to consumers the meaning of quality of care. Med Care Res Rev. 2010;67(3):275-93.

46. Hibbard J, Sofaer S. Best Practices in Public Reporting No. 1: How to Effectively Present Health Care Performance Data to Consumers. AHRQ Publication No. 10-0082-EF. Rockville, MD: Agency for Healthcare Research and Quality, June 2010.

About the Authors

Mark W. Friedberg, MD, MPP is a Senior Natural Scientist at the RAND Corporation and an Assistant Professor of Medicine, part time, at Harvard Medical School and Brigham and Women's Hospital.

Peter J. Pronovost, MD, PhD, FCCM is the Johns Hopkins Medicine Senior Vice President for Patient Safety and Quality and Director of the Armstrong Institute for Patient Safety and Quality.

David M. Shahian, MD is Professor of Surgery, Harvard Medical School; Vice-President, Center for Quality and Safety, Massachusetts General Hospital; Chair of the Society of Thoracic Surgeons Quality Measurement Task Force; and a member of the National Quality Forum Board of Directors.

Dana Gelb Safran, Sc.D. is Senior Vice President, Performance Measurement & Improvement, Blue Cross Blue Shield of Massachusetts and Associate Professor of Medicine, Tufts University School of Medicine.

Karl Y. Bilimoria, MD, MS is a surgical oncologist, Director of the Surgical Outcomes and Quality Improvement Center, Vice Chair for Quality in the Department of Surgery at Northwestern University's Feinberg School of Medicine, and the Director of Surgical Quality for Northwestern Memorial Hospital.

Marc N. Elliott, PhD is a Senior Principal Researcher at the RAND Corporation and holds the RAND Distinguished Chair in Statistics.

Cheryl L. Damberg, PhD is a Senior Principal Researcher at the RAND Corporation and holds the RAND Distinguished Chair in Health Care Payment Policy.

Justin B. Dimick, MD, MPH is Professor of Surgery and Health Management & Policy; Director, Center for Healthcare Outcomes & Policy; Chief, Division of Minimally Invasive Surgery; and Associate Chair for Strategy and Finance, Department of Surgery, University of Michigan.

Alan M. Zaslavsky, PhD is a Professor of Health Care Policy at Harvard Medical School.

About This Report

None of the authors of this document received any form of remuneration in exchange for his or her contributions to it. RAND Health, a division of the RAND Corporation, supported the editing and production costs for this RAND Perspective using funds derived from the generosity of RAND's donors and the fees earned on client-funded research. A profile of RAND Health, abstracts of its publications, and ordering information can be found at www.rand.org/health.

Conflict of Interest Statement

Dr. Pronovost is an employee of Johns Hopkins University School of Medicine, where he is Senior Vice President for Patient Safety and Quality and Director of the Armstrong Institute for Patient Safety and Quality. In these roles, Dr. Pronovost reviewed internal performance data for a Johns Hopkins surgeon who had a high "Adjusted Complication Rate" on the ProPublica *Surgeon Scorecard* and met with this surgeon and his department director. The ProPublica article "Making the Cut: Why Choosing the Right Surgeon Matters Even More Than You Know,"[21] which accompanied the publication of the *Surgeon Scorecard*, named this surgeon. In 2014, ProPublica requested and received comments from Dr. Pronovost on the goals and methods of the *Surgeon Scorecard* prior to its publication.

Dr. Shahian, Dr. Bilimoria, and Dr. Dimick are surgeons with special interest and expertise in developing and reporting methodologically credible, peer-reviewed performance metrics. None of the authors of this paper has an "Adjusted Complication Rate" reported in the current ProPublica *Surgeon Scorecard*. Dr. Shahian is a member of the Board of Directors and the Executive Committee of the National Quality Forum. Dr. Dimick is a co-founder of ArborMetrix, Inc., a company that provides software for measuring hospital quality and efficiency.

No other author has a conflict of interest.

Limited Print and Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited. Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please visit www.rand.org/pubs/permissions.html.

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. **RAND**® is a registered trademark.

For more information on this publication, visit www.rand.org/t/PE170.

www.rand.org

© Copyright 2015 RAND Corporation

PE-170 (2015)