RAND CORPORATION

Perspective

Expert insights on a timely policy issue

Response to ProPublica's Rebuttal of Our Critique of the Surgeon Scorecard

By Mark W. Friedberg, Karl Y. Bilimoria, Peter J. Pronovost, David M. Shahian, Cheryl L. Damberg, and Alan M. Zaslavsky

In the summer of 2015, ProPublica published its *Surgeon Scorecard*, which displays "Adjusted Complication Rates" for individual, named surgeons for eight surgical procedures performed in hospitals. In September 2015, RAND released a critique of the *Scorecard* authored by a group of heath policy researchers from RAND and other institutions.[1] On October 7, 2015, ProPublica published a rebuttal of our critique of the *Surgeon Scorecard*.[2] However, our methodological concerns regarding the *Scorecard* remain unaddressed. Therefore, we continue to advise potential users of the *Scorecard*, as it is currently constructed, not to consider it a valid or reliable predictor of the health outcomes any individual surgeon is likely to provide.

In our critique, we explained five major methodological shortcomings in the *Scorecard*. Here, we revisit them point by point, summarize ProPublica's rebuttals, and explain why each rebuttal does not resolve the corresponding methodological problem.

POINT 1 FROM OUR INITIAL CRITIQUE

The "Adjusted Complication Rates" reported in the *Scorecard* are not actually complication rates.

Moreover, there is no credible scientific evidence that differences between surgeons' "Adjusted Complication Rates" are associated with differences in short- or long-term patient health outcomes following surgery, and there are substantial reasons to believe such associations are weak to nonexistent. Two methodological concerns underlie this point. First, ProPublica's "Adjusted Complication Rates" are primarily 30-day readmission rates for conditions plausibly related to the index surgery. Because these "Adjusted Complication Rates" exclude complications occurring during the index admission, they are not accurate estimates of overall operative complication rates, even in the short term. Second, there is no credible empirical evidence that surgeon-to-surgeon differences in the readmissions counted in the "Adjusted Complication Rates" are actually associated with differences in short- or long-term

rates of underlying complications (rather than differences in propensity to readmit a given patient).

PROPUBLICA'S REBUTTAL (PART 1)

Regarding our first concern, ProPublica objects to our citation of a study finding that that 67.1 percent of 30-day surgical complications occur during the index admission because this figure includes a wider variety of procedures than those reported in the *Scorecard*.[3] ProPublica's rebuttal also states, based on a comparison to procedures such as breast lumpectomies (most of which occur in the outpatient setting), "that for low-risk procedures comparable to those included in *Scorecard*, more than 90 percent of complications are captured by tracking what happens after a patient leaves the hospital."

OUR RESPONSE (PART 1)

It is true that the 67.1-percent figure includes a wider variety of procedures than those reported in the *Scorecard*, both inpatient (such as those in the *Scorecard*) and outpatient (which are not included in the *Scorecard*). To see whether the 67.1-percent figure was an inaccurate assessment, we queried the full 2011–2014 American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database for seven of the eight specific elective procedures reported in the *Scorecard* (analyzing data from more than 57,000 operations), applying nearly the same inclusion and exclusion criteria used in the *Scorecard*.[4] We found that among these inpatient procedures reported in the *Scorecard*, 88 percent of 30-day complications occurred during the index admission, ranging from 64 percent (for laparoscopic cholecystectomy) to 90 percent (for total hip replacement). This finding suggests that the *Scorecard* fails to detect in its "Adjusted Complication Rates" the majority of 30-day complications for all seven of these procedures. We are unaware of any empirical data that supports ProPublica's assertion that the *Scorecard* only misses 10 percent of such complications.

Moreover, we were unable to find any scientifically credible empirical evidence that surgeons with higher post-discharge "Adjusted Complication Rates" also have higher overall complication rates within 30 days. There is reason to believe this association is likely to be weak to nonexistent. To illustrate one aspect of the problem, surgeons who achieve earlier discharge for their patients (which is, all other things equal, a good thing) will tend to have lower index admission complication rates but higher "Adjusted Complication Rates" than surgeons who discharge their patients later, even if both groups of surgeons have the exact same underlying true rates of complications. This is simply because complications that would have manifested during the index admission for a slow-to-discharge surgeon instead manifest after discharge for those surgeons who discharge sooner.

PROPUBLICA'S REBUTTAL (PART 2)

Regarding our second concern (the lack of evidence that readmissions counted in the "Adjusted Complication Rates" are actually associated with differences in the rates of underlying complications), ProPublica's rebuttal contains no direct counterargument. They do state that "RAND's authors argue that *Scorecard* is deficient because it doesn't rely on clinical data—detailed information from patients' medical charts, notes and internal hospital record-keeping systems. But such information is not publicly available, and there is no national database of clinical information, public or private. We didn't neglect to use it—it doesn't exist. When the authors argue for reporting that relies on clinical data, they are really arguing for no national reporting at all."

OUR RESPONSE (PART 2)

ProPublica's rebuttal does not address our methodological concern and, unfortunately, both mischaracterizes our critique and misrepresents the data requirements for measure validation. Nowhere in our critique did we suggest that ProPublica should replace claims data with clinical data, or that public reporting of claims-based measures should be abandoned.

What we *did* suggest—and continue to recommend—is that ProPublica perform credible validation of its claims-based "Adjusted Complication Rates" using a comparison to *a representative sample* of clinical data. As others have demonstrated, existing registry databases can be used to validate claims-based performance measures for national reporting.[5]

As an alternative to empirically validating its "Adjusted Complication Rates" as measures of surgical complications, ProPublica could rename the measures reported in the *Scorecard* (i.e., step back from the unsubstantiated claim that "Adjusted Complication Rates" are accurate indicators of the relative complication rates patients will experience by choosing one surgeon over another).

POINT 2 FROM OUR INITIAL CRITIQUE

As currently constructed, the *Scorecard* masks hospital-to-hospital performance differences, thereby invalidating comparisons between surgeons in different hospitals.

This shortcoming is a consequence of ProPublica's methodological choice to set hospital random effects equal to zero in calculating the numerical "Adjusted Complication Rates" displayed in the *Scorecard*.

PROPUBLICA'S REBUTTAL (PART 1)

ProPublica writes "Contrary to what the RAND authors assert, there is a perfectly good reason for isolating a surgeon's performance from a hospital. It is to level the playing field among surgeons. We held constant the effects of the hospital (and the patient pool) to assure, as much as possible, that surgeons' scores were not dragged down by things beyond their control. We believe this is the best way to be fair to doctors who are, after all, the ones whose names appear in our database."

OUR RESPONSE (PART 1)

The rationale given by ProPublica in its rebuttal for not taking hospital random effects into account in the *Scorecard*'s "Adjusted Complication Rates" is false. As we explained in our original critique, hierarchical models such as the ones used to create the *Scorecard* cannot distinguish between performance clustering due to intrinsic hospital factors (e.g., quality of nursing staff) and performance clustering due to intentional or unintentional recruitment of surgeons whose *individual* skill levels are higher or lower than average. In other words, both hospital factors (which might be called "beyond surgeons' control") and recruitment (which is an aggregation of surgeons' inherent skills as individuals) contribute to hospital random effect estimates. As a consequence, ProPublica's assertion that setting hospital random effects equal to zero enables fair comparison between surgeons in different hospitals would be true only if there were negligible surgeon recruitment based on ability (accepting, for sake of argument, ProPublica's implicit and, as discussed above, unproven assumption that surgeon ability is associated with "Adjusted Complication Rates"). Such a zero-recruitment scenario is extremely unlikely to reflect reality. Therefore, setting hospital random effects equal to zero will penalize truly excellent surgeons (and reward truly poor ones) who happen

to be recruited to the same hospital based on their good (or bad) surgical abilities as individuals. Thus, there is no sound rationale for ProPublica's methodological choice, which cannot "level the playing field" among surgeons in different hospitals without, to an unknown extent, also unintentionally leveling their intrinsic individual abilities.

PROPUBLICA'S REBUTTAL (PART 2)

In the "near me" view within the *Scorecard*, hospitals are "sorted by the surgeon with the lowest adjusted rate of complications at each hospital, along with a measure representing the combined performance of surgeons and hospitals for these procedures." In other words, in this data view only, ProPublica includes (rather than ignores) hospital random effects estimates in the "Adjusted Complication Rates" displayed.

OUR RESPONSE (PART 2)

ProPublica's response implicitly assumes that report users will restrict their between-hospital comparisons to the "near me" view of the *Scorecard*, never making a comparison between hospitals by looking at the within-hospital scores of individual surgeons (which do have hospital random effects set to zero) across two or more hospitals. To our knowledge, there is no empirical basis for such an assumption. Indeed, local news media outlets have facilitated comparisons between surgeons in different hospitals using "Adjusted Complication Rates" for which the hospital random effects were set to zero, thereby misleading their readers in the manner we have described.[6]

Reporting relatively more-accurate performance data in one part of a report but retaining less-accurate and potentially misleading performance data in another is without precedent, is a major divergence from best practices in public reporting, is not a strength of the *Scorecard*, and is not equivalent to correcting the underlying methodological error—which is straightforward and possible to do in the current *Scorecard* by including hospital random effects estimates in all reported "Adjusted Complication Rates."

POINT 3 FROM OUR INITIAL CRITIQUE

The accuracy of the attribution of performance data to specific surgeons in the *Scorecard* is questionable.

This concern stems from our observation of obvious wrong-specialty misattributions of surgical cases to inapplicable physicians in the initial release of the *Scorecard*, as well as prior research finding high rates of disagreement between Part A and Part B Medicare claims for surgical procedures.[7]

PROPUBLICA'S REBUTTAL

ProPublica reports that obvious misattributions (e.g., to non-surgeons or even providers who are not physicians) have been corrected in the *Scorecard*.

OUR RESPONSE

ProPublica's rebuttal does not address the concern articulated in our critique, which is that the rate of case misattribution between surgeons in the *Scorecard* is unknown. There is no substitute for due diligence: confirming the accuracy of the source data by comparison to a gold standard in a credible sample of surgeons and hospitals.

We also note that despite ProPublica's statement that such errors have been corrected, at the time of this writing, two inapplicable

physicians (an interventional cardiologist and a heart surgeon—both of whom we mentioned in our critique) are still listed in the *Scorecard* as having performed hip replacements at Massachusetts General Hospital. It is unlikely that these are the only remaining inaccuracies, even among those that should be detected readily by examining physician specialty information.

POINT 4 FROM OUR INITIAL CRITIQUE

The adequacy of the *Scorecard*'s case-mix adjustment is questionable.

This concern stems from the lack of credible empirical evidence that case-mix adjustment was valid.

PROPUBLICA'S REBUTTAL

The rebuttal states that the procedures reported in the *Scorecard* are usually elective, repeats the *Scorecard*'s case-mix adjustment methods, and offers these statements: "Many surgeons in *Scorecard* have good scores even though they work in areas known for unhealthy populations. Likewise, there are many surgeons with good scores at hospitals known for taking on some of the toughest cases."

OUR RESPONSE

ProPublica's rebuttal does not address our methodological concern. In the absence of scientifically credible validation of the case-mix methodology, as detailed in our original critique, there is no empirical evidence that the case-mix adjustment in the *Scorecard* yields fair comparisons between surgeons. Our critique cites examples of how such validation can be performed.

POINT 5 FROM OUR INITIAL CRITIQUE

The *Scorecard* appears to have poor measurement reliability and therefore randomly misclassifies the performance of many surgeons.

This concern is based on the confidence intervals presented in the *Scorecard*, which are wide relative to the distribution of point estimates.

PROPUBLICA'S REBUTTAL (PART 1)

ProPublica writes "The 'misclassification' alleged by the RAND authors does not exist because no classification occurs in the first place."

OUR RESPONSE (PART 1)

By standard definitions of the word "classification" (and certainly the definition underlying the methodological literature on performance misclassification), the *Scorecard* does classify each surgeon's "Adjusted Complication Rate" as low, medium, or high—with corresponding green, yellow, and red colors to accentuate the differences between categories. This classification system is ProPublica's basis for applying red exclamation points to hospitals with one or more surgeons whose "Adjusted Complication Rates" are in the "high" category for a procedure.

PROPUBLICA'S REBUTTAL (PART 2)

ProPublica writes "The RAND authors argue that most users are 'likely' not sophisticated enough to understand the confidence intervals around surgeon complication rates in *Scorecard*. They offer no direct evidence."

OUR RESPONSE (PART 2)

ProPublica's rebuttal does not address our fundamental concern regarding reliability, which is that the *Scorecard* has no minimum

reliability requirement at all. In fact, there is no evidence in any of the documentation on the *Scorecard* published to date that ProPublica has calculated reliability estimates. This is a fundamental step in performance reporting, and one that well-constructed reports follow (such as those reporting HCAHPS scores). Our original critique contains more detail on the importance of measurement reliability in performance reporting.

ProPublica's use of confidence intervals is not a substitute for requiring a minimum reliability threshold. Though ProPublica asserts that *Scorecard* users can interpret confidence intervals, this assertion conflicts with years of prior empirical research (Sofaer and Hibbard,[8] which we cited in our critique, is a good introduction to the literature on this topic). While it is possible that ProPublica is right and that prior research is wrong, ProPublica has not publically disclosed the results of usability testing that follows credible scientific standards.

Beyond the ability of a lay audience to interpret their meaning, confidence intervals are easily overshadowed by the corresponding point estimates, no matter how unreliable the estimate. Indeed, other media outlets have reported point estimates and performance classifications from the *Scorecard* without mentioning the statistical uncertainty accompanying these estimates.[9] Having a minimum reliability criterion and abiding by it would limit the extent to which such potentially misleading communication could occur.

Empirical Research and the Burden of Proof

Finally, we note that ProPublica's description of our methodological concerns as "conjecture" is true, at least insofar as "conjecture" means "unproven hypothesis." At this time we cannot rule out with absolute certainty the possibility—remote though it may be—that despite our concerns, the *Scorecard* is a valid and reliable indicator of the health outcomes a given surgeon is likely to deliver. After all, anything is *possible* until proven otherwise.

But developing a new performance measure is a research endeavor, and to believe that a scientific investigator's new, untested assertions are true until they are disproven is to deviate fundamentally from usual practice in empirical research. As a general principle, researchers do not automatically consider new hypotheses to be true until competing hypotheses are shown to be false (or at least unlikely to be true, using credible methods), leaving the new hypothesis as the only possible explanation (or most likely explanation) for the observed data. It is possible that the standards of evidence are different in data journalism—that a data journalist's untested interpretation of his or her analyses must be considered true, until contradictory evidence is produced. If so, these evidentiary standards would be worth explaining in more detail.

Stated another way, as of this writing, the following assertions implicit in the *Scorecard* are unsupported by credible scientific evidence (i.e., are "conjecture," following ProPublica's nomenclature). Yet each of these assertions is essential to establishing the credibility of the ProPublica report card:

- **Unsupported assertion #1:** The "Adjusted Complication Rates" reported in the ProPublica Surgeon *Scorecard*—if measured without error and with perfect case-mix adjustment—are valid predictors of the relative health outcomes produced by individual surgeons.
- **Unsupported assertion #2:** The source data used to calculate the "Adjusted Complication Rates" for a given surgeon reflect the care delivered by that surgeon, rather than care delivered by other surgeons.
- **Unsupported assertion #3:** The "Adjusted Complication Rates" reported in the ProPublica Surgeon *Scorecard* are adequately risk-adjusted.

- **Unsupported assertion #4:** The reliability of the "Adjusted Complication Rates" reported in the *Scorecard* is acceptable and understandable to potential report users.

Until ProPublica or other interested parties perform the analyses and due diligence usually conducted to validate a new measure such as the *Scorecard*'s "Adjusted Complication Rates," check the accuracy of the source data, validate the risk adjustment, and calculate measurement reliability, there is no way to know whether the data presented in the *Scorecard* are valid or reliable predictors of anything of value to patients, providers, or other stakeholders. As we explained in our original critique, it is entirely possible for an invalid, unreliable performance report to harm patients, both in the short and long term—i.e., to be worse than nothing at all. And to assume the validity and reliability of a brand-new measure in light of the problems we have raised is to take a remarkable leap of faith, without grounding in best practices for performance reporting.

Endnotes

[1] Friedberg, Mark W., Peter J. Pronovost, David M. Shahian, Dana Gelb Safran, Karl Y. Bilimoria, Marc N. Elliott, Cheryl L. Damberg, Justin B. Dimick, and Alan M. Zaslavsky, *A Methodological Critique of the ProPublica Surgeon Scorecard*, Santa Monica, Calif.: RAND Corporation, PE-170, 2015. As of December 1, 2015: http://www.rand.org/pubs/perspectives/PE170.html

[2] Engelberg, Stephen, and Olga Pierce, "Our Rebuttal to RAND's Critique of Surgeon Scorecard," *ProPublica*, October 7, 2015. As of December 1, 2015: https://www.propublica.org/article/our-rebuttal-to-rands-critique-of-surgeon-scorecard

[3] Bilimoria, Karl Y., Mark E. Cohen, Angela M. Ingraham, David J. Bentrem, et al., "Effect of Postdischarge Morbidity and Mortality on Comparisons of Hospital Surgical Quality," *Annals of Surgery*, Vol. 252, No. 1, 2010, pp. 183–190.

[4] We were able to reliably exclude revisional surgeries for all procedures given coding limitations, thus they are included in our sample. Also, the NSQIP included only 18 cases of inpatient transurethral prostatectomy among patients aged 65 or older, precluding meaningful analysis of this procedure (though we note that among these 18 cases, 100 percent of observed complications occurred during the index admission).

[5] Krumholz, Harlan M., Yun Wang, Jennifer A. Mattera, Yongfei Wang, Lein Fang Han, Melvin J. Ingber, Sheila Roman, and Sharon-Lise T. Normand, "An Administrative Claims Model Suitable for Profiling Hospital Performance Based on 30-Day Mortality Rates Among Patients with an Acute Myocardial Infarction," *Circulation*, Vol. 113, No. 13, April 4, 2006a, pp. 1683–1692. As of December 2, 2014: http://www.ncbi.nlm.nih.gov/pubmed/16549637

[6] For example, see Traci Moyer, "Investigation: Valley Surgeons, A Cut Above Peers," *News Leader*, July 17, 2015 (as of December 1, 2015: http://www.newsleader.com/story/news/local/2015/07/17/investigation-propublica-surgeon-scorecard-complications-augusta-health-rockingham-uva-mary-washington/30113745/); and Ben Sutherly, "Surgeon Ratings Get Mixed Marks," *Columbus Dispatch*, July 19, 2015 (as of December 1, 2015: http://www.dispatch.com/content/stories/local/2015/07/19/surgeon-ratings-get-mixed-marks.html).

[7] Dowd, Bryan, Robert Kane, Shriram Parashuram, Tami Swenson, and Robert F. Coulam, *Alternative Approaches to Measuring Physician Resource Use: Final Report,* Centers for Medicare and Medicaid Services, April 9, 2012. As of December 1, 2015: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Reports/Research-Reports-Items/Alternative-Approaches-to-Measuring-Physician-Resource-Use.html

[8] Sofaer, Shoshanna, and Judith Hibbard, *Best Practices in Public Reporting No. 3: How to Maximize Public Awareness and Use of Comparative Quality Reports Through Effective Promotion and Dissemination Strategies*, Rockville, Md.: Agency for Healthcare Research and Quality, June 2010. As of December 2, 2015: http://archive.ahrq.gov/professionals/quality-patient-safety/quality-resources/tools/pubrptguide3/pubrptguide3.pdf

[9] See, for example, Moyer, 2015; and Sutherly, 2015.

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," 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. Bilimoria and Dr. Shahian 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.

No other author has a conflict of interest.

About This Perspective

The authors gratefully acknowledge Dr. Kristen Ban for analyzing NSQIP registry data. 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.

For more information on this publication, visit www.rand.org/t/pe170z1.

www.rand.org