The NEW ENGLAND JOURNAL of MEDICINE

Perspective

OCTOBER 8, 2015

Scoring No Goal — Further Adventures in Transparency

Lisa Rosenbaum, M.D.

One Monday morning, rounding on a patient who needed relatively urgent coronary-artery bypass surgery, a newly appointed cardiologist in New York asked the team to call a surgical consult (some details have been changed to protect those involved). "We can't call today," the cardiology fellow explained patiently. "Dr. X. is taking consults. He wouldn't touch our patient with a 10-foot pole." The fellow scrutinized the call schedule. "The only surgeon who might take him isn't on until Wednesday."

In New York, one of a handful of states where outcomes of cardiac surgery and percutaneous coronary intervention (PCI) have been publicly reported for some years, such situations are not uncommon. Although the aim of report cards was to motivate proceduralists to improve quality, they seem instead to have motivated avoidance of the sickest patients. Studies of PCI suggest that such patients — those with heart failure or cardiogenic shock, for example — are less likely than healthier ones to receive revascularization.[1] Whereas one analysis found no attendant mortality difference, another recent study comparing patients with acute myocardial infarction in states with and states without public reporting showed higher mortality in public-reporting states — a difference driven by deaths among critically ill patients in whom revascularization was not undertaken.[2] Though debate therefore continues over the risk–benefit balance of public reporting, no one denies the need for quality improvement and sound information to help patients choose a provider. The problem is that it's not clear that such scorecards achieve either.

In this age of ever-increasing transparency, it's not surprising that such scoring efforts nevertheless continue. The latest, a "surgeon scorecard," was released in July by ProPublica, which produces "investigative journalism in the public interest." ProPublica is known for data-transparency initiatives such as "Dollars for Docs," which provides a searchable database of industry payments to individual physicians. The surgeon scorecard, based on Medicare claims data for 2009 through 2013, analyzes the complication rates of 17,000 surgeons for eight elective procedures.[3] Inpatient mortality and readmissions data were used to identify instances of possible harm; then cases with a diagnostic code for a surgery-related complication

were used to calculate complication rates. Risk adjustment factored in coexisting conditions such as obesity and diabetes, and a mixed-effects model was used to control for variables such as hospital quality. By using these logistic-regression techniques to generate a categorical score, the organization leapt beyond presenting factual information ("Your doctor received $1,000 from industry last year") to presenting information imbued with ostensible statistical meaning ("Your surgeon's complication rate is higher than average"). ProPublica has thus migrated from the realm of data journalism to scientific analysis.

The medical community's reaction was swift, incisive, and highly critical, raising two chief concerns about the data. One is that insurance claims data are notoriously inaccurate, particularly when it comes to assessing surgical complications. As Dartmouth–Hitchcock surgeon and quality expert John Birkmeyer explains, "They used administrative data to measure the one component of quality for which administrative data have been shown to be most flawed." Let's say a patient is readmitted after a prostatectomy and the diagnostic code is for deep-vein thrombosis (DVT). Aside from the fact that codes are often inaccurate, is it the surgeon's fault if the patient has a DVT? Patients don't always take prescribed anticoagulants, and even those who do, particularly those with underlying cancer, sometimes get DVTs. Without the granularity of a chart review, assigning responsibility is impossible.

Even good data, however, can't overcome the second limitation, which holds for any attempt to rank individuals: poor reliability. There were too few surgeries for any given surgeon to generate statistically meaningful results. As Justin Dimick, a surgeon and outcomes researcher at the University of Michigan, explains, even when you aggregate outcomes at the hospital level, numbers are generally too small. Dimick's research group studied quality indicators for Michigan colorectal surgeons and found only one such surgeon in the entire state whose volume was high enough to permit statistically meaningful assessment of outcomes.[4] ProPublica encountered the same hurdle, which resulted in wide and overlapping confidence intervals for most surgeons . . . who were nevertheless assigned a score. Mark Friedberg, a RAND scientist, captured the problem aptly in a tweet: "With confidence intervals like these, who needs enemies?"

Though Friedberg was referring to the egregious statistics, his observation seems equally relevant to the disingenuous tone of the effort — the professed commitment to protecting patients and helping us improve, as a guise for frank fear mongering. Before publishing the scorecard, ProPublica released a promotional video, set to melodramatic music, alerting the public about the "400,000 patients harmed in our hospitals," as the word "avoidable" flashes across the screen. The sensationalism continued in the feature article accompanying the scorecard, which names particular surgeons, including a Johns Hopkins urologist with a high complication rate — "surprising," says the article, given that Hopkins "is home to Peter Pronovost, a foremost leader in the patient safety movement."[3] When it's revealed that Pronovost was the only expert consulted who was critical of ProPublica's methods, the implication is that his criticisms reflect institutional fidelity rather than legitimate concerns about the data analysis.

Soon, however, the lone critic was joined by others who delineated some important problems. For starters, procedural quality is about more than complication rates. As University of Pittsburgh urologist Benjamin Davies pointed out in a Forbes.com blog post, for instance, prostatectomy aims to remove cancer and preserve urinary and erectile function; grading it solely on whether the patient develops a "digestive system complication" overlooks what matters to many patients. "Would you trade a bad bout of constipation for an erection?" he asks. "I would."

Highlighting the false assumption that report cards protect patients by rooting out bad doctors, University of Pennsylvania radiologist Saurabh Jha, writing at the Health Care Blog (www.thehealthcareblog.com), decried the ability of the scorecard to destroy the reputations and careers of excellent doctors. In an invented scenario that Jha envisions as one likely consequence, a surgeon who cherry-picks patients handily outscores the safety-net-hospital star to whom everyone sends their sickest patients. When the latter surgeon is revealed to have the city's highest complication rates and the referrals stop coming, his professional pride — a concept, notes Jha, that "eludes some health economists" — is shattered, to the detriment of those sick patients.

Despite such shortcomings, scorecards will continue to be

developed, and ProPublica has at least triggered an important discussion about how to do it right.

Robert Yeh, a Massachusetts General Hospital cardiologist who studies public reporting, offers several insights from the cardiology experience. First, says Yeh, claims data have quite limited utility for ranking individuals. Massachusetts, for instance, depends on detailed registry data collected at the point of care for the sole purpose of prospectively tracking quality. Second, critical variables used to stratify risk, such as the presence of shock or infection, must be adjudicated by experts who meet regularly and review charts, thereby removing physicians' incentive to make patients "look" sicker than they are. Finally, the process itself needs to undergo constant "quality improvement." Given the continued avoidance of the highest-risk patients, for example, Massachusetts now excludes from individual physician scores patients who fall into "compassionate use" or "exceptional risk" categories.

Yeh's most critical point, however, is more philosophical than methodologic: "Transparency becomes increasingly necessary in an environment with low trust." The widespread perception that we as a profession are failing to "right ourselves" lends these transparency efforts a triumphant aura of progress. Health policy expert Ashish Jha, in defending the scorecard on a Harvard School of Public Health blog, suggests that the perception of the profession's inaction is not misguided. "We could do much better," Jha argues, "but we have chosen not to." The National Surgical Quality Improvement Program, for instance, collects surgical quality information; 600 hospitals use it, 3000 don't. Jha argues that such data should be made publicly available and that ProPublica's efforts establish a new standard for us to improve on. But even if the scorecard compels similar efforts from within medicine that help to solve our image problem, we still face the limitations of even rigorously collected registry data: small numbers and dichotomous variables that cannot capture true expertise. The key question, then, is less about transparency with regard to quality than it is about what constitutes quality in the first place.

Recently, a friend of mine in another city asked me to recommend a surgeon for a routine surgical procedure. I did what I assume most of us would do: I called a physician-friend in that city and asked him who was good. The notion that peer impressions are the best proxy for quality underlies a common lament about scorecards: "The best surgeon I know has the highest complication rates." But that notion has also driven some of the most creative work in assessing surgical quality.

Birkmeyer and colleagues had peer reviewers rate videos of bariatric surgeons performing gastric bypass procedures, and then assessed the relationship between these ratings and complication rates.[5] They found that surgical skill varied widely, and greater technical skill as judged by peers predicted better outcomes. Videos of high- and low-performing surgeons operating were also published, and most intriguing to Birkmeyer was how easily technical skill could be assessed by viewers — and not only physician-viewers. "Within 5 seconds," he told me, "every layperson could tell who was skilled and who was not."

Birkmeyer believes such video-based assessment will be critical to future quality and transparency efforts, at least in surgery. As part of a broader effort to establish standards across hospitals in the Dartmouth–Hitchcock system, he is pushing surgeons in some specialties to use videos for credentialing and peer review; the videos would later be made publicly available online. Though we have much to learn about how to translate such approaches for nonprocedural fields and about whether quality assessment leads to improvement, Birkmeyer is pragmatic: "We can't hold ourselves hostage to empirical perfection."

Empirical imperfection, however, is far different from pseudo-empiricism. The real danger of scorecards like ProPublica's lies not in their failure to objectively capture quality but in their pretense that they succeed. Though some observers see such scorecards as superior to online ratings like those on Yelp, I disagree. Few people know how to interpret overlapping confidence intervals like ProPublica's, whereas Yelp fully embraces its inherent subjectivity. Indeed, when I suggested the reputable surgeon to my friend, she promptly discovered a negative review describing his insensitivity when counseling a patient to lose weight. She chose another surgeon. Though I initially balked, I soon realized that we were both doing the same thing: relying on our peers to assess the qualities we value.

The irony in hailing the scorecard as a victory for transparency is that its purported objectivity obscures its methodologic limita-

tions and the complexity of quality itself. No amount of transparency can overcome the fact that, when it comes to what we value, we don't all see eye to eye. The real promise of transparency, then, lies in finding better ways to let our patients see what we see.

Disclosure forms provided by the author are available with the full text of this article at NEJM.org

Dr. Rosenbaum is a national correspondent for the *Journal*.

This article was published on September 2, 2015, at NEJM.org.

1. Joynt KE, Blumenthal DM, Orav EJ, Resnic FS, Jha AK. Association of public reporting for percutaneous coronary intervention with utilization and outcomes among Medicare beneficiaries with acute myocardial infarction. JAMA 2012;308:1460-8.
2. Waldo SW, McCabe JM, O'Brien C, Kennedy KF, Joynt KE, Yeh RW. Association between public reporting of outcomes with procedural management and mortality for patients with acute myocardial infarction. J Am Coll Cardiol 2015;65:1119-26.
3. ProPublica. Surgeon scorecard (https://projects.propublica.org/surgeons).
4. Shih T, Cole AI, Al-Attar PM, et al. Reliability of surgeon-specific reporting of complications after colectomy. Ann Surg 2015;261:920-5.
5. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434-42.

DOI: 10.1056/NEJMp1510094

Pledging to Eliminate Low-Volume Surgery

David R. Urbach, M.D.

On May 18, 2015, leaders at three hospital systems — Dartmouth–Hitchcock Medical Center, the Johns Hopkins Hospital and Health System, and the University of Michigan Health System — publicly announced a "Take the Volume Pledge" campaign to prevent certain surgical procedures from being performed by their surgeons who perform relatively few of them or at their hospitals where relatively few such procedures are performed. The Pledge, promoted by longtime advocates of quality improvement such as John Birkmeyer and Peter Pronovost, challenges other large health systems to join them in restricting the performance of 10 surgical procedures — including gastrointestinal, cardiovascular, and joint-replacement surgeries — to hospitals and surgeons who perform more than a minimum number. The annual volume thresholds range from 10 per hospital and 5 per surgeon for carotid stenting to 50 per hospital and 25 per surgeon for hip and knee replacement.

The reaction of surgeons to the Pledge, which was widely promoted in *U.S. News & World Report* as part of its Best Hospitals for Common Care rankings, was predictably hostile — and completely out of proportion to the modest ambition of the Pledge. Of all the possible approaches to restricting surgical care to high-volume hospitals, perhaps the least controversial ought to be a decision by a large metropolitan academic hospital system that its most complex elective surgery should be performed by the providers and hospitals that do the most of a given procedure. (The Pledge is silent on the question of performing complex surgery in independent small and rural hospitals.) If volume-based distribution of surgery cannot be accomplished in this context, then it's probably not going to happen anywhere.

Proponents of the Pledge have presumably calculated that starting with such an easily achievable policy could be the thin end of the wedge for broader efforts to centralize complex surgery. Nevertheless, a discussion board of the American College of Surgeons quickly filled with dozens of postings, almost all bristling at the idea of external organizations imposing volume standards on surgery, arguing instead for quality-based standards, and taking particular offense at the portrayal of low-volume surgeons as hobbyists who are motivated by professional autonomy and pride to continue performing rare procedures despite the clinical and economic consequences.[1]

For anyone wondering why we are still discussing surgical volume 36 years after Harold Luft pointed out the relationship between higher surgical volume and lower postoperative mortality,[2] and why even limited initiatives such as the Pledge elicit such controversy, it is useful to reflect on how we got here and what it means for the prospects of quality improvement in surgery.

There is no doubt that the outcomes of elective surgery — as measured by postoperative mortality, complications, or a wide array of other measures — are better when an operation is done by a surgeon or in a hospital with a high procedure volume. The volume–outcome effect is a remarkably consistent finding in studies of surgical services, and it applies not only to surgery, but also to many types of nonsurgical hospital-based care, such as treatment of congestive heart failure and chronic obstructive