Chapter 3
Measuring Surgical Quality

Andrew M. Ibrahim and Justin B. Dimick

Abstract Improving surgical care requires an in depth understanding of how to measure quality. We are currently witnessing an unprecedented level of investment from payers, policy makers, patient advocacy groups and professional societies to measure the quality of care surgeons provide. Despite the widespread interest in measuring quality, there is little consensus how it should be done. Payers and regulators often target processes of care (e.g. appropriate use of preoperative antibiotics), while surgeons tend to focus on outcomes that are seen as the "bottom line" (e.g. 30-day post-operative mortality rates.) Most recently, numerous stakeholders are advocating for the use of patient reported information (e.g. "How did this operation affect your daily living?")

Abbreviation

PROs Patient Reported Outcomes

Improving surgical care requires an in depth understanding of how to measure quality. We are currently witnessing an unprecedented level of investment from payers, policy makers, patient advocacy groups and professional societies to measure the quality of care surgeons provide. Despite the widespread interest in measuring quality, there is little consensus how it should be done. Payers and regulators often target

A.M. Ibrahim, MD, MSc
Institute for Healthcare Policy & Innovation, University of Michigan, Ann Arbor, MI, USA

Department of Surgery at Michigan, University Hospitals-Case Medical Center No 2, Cleveland, OH, USA

J.B. Dimick, MD, MPH (✉)
Division of Minimally Invasive Surgery, Center for Healthcare Outcomes and Policy, Ann Arbor, MI, USA

Department of Surgery, University of Michigan, Ann Arbor, MI, USA
e-mail: jdimick@umich.edu

R.R. Kelz, S.L. Wong (eds.), *Surgical Quality Improvement*,
Success in Academic Surgery, DOI 10.1007/978-3-319-23356-7_3

processes of care (e.g. appropriate use of preoperative antibiotics), while surgeons tend to focus on outcomes that are seen as the "bottom line" (e.g. 30-day post-operative mortality rates.) Most recently, numerous stakeholders are advocating for the use of patient reported information (e.g. "How did this operation affect your daily living?")

Each of these strategies has benefits and drawbacks that make each different approach appropriate in a specific context. In addition to the goals specified in each context, there are important statistical limitations that can constrain our ability to meaningfully measure quality. This chapter begins with an overview of common measurement approaches and introduces emerging approaches as well. We then review relevant statistical concepts that inform how to choose between each measurement approach.

What Should We Measure? The Structure, Process, Outcomes Framework

The most common framework to measure quality is the "Structure, Process, Outcomes" model described by Donabedian in 1998 (Donabedian 1988). Each component of the model is described below with the benefits and drawbacks of using them in a surgical context (Table 3.1).

Table 3.1 Current approaches to measuring quality in surgery: structure, process outcomes

Type of measure	Example	Benefits	Drawbacks
Structure			
	Hospital Volume	Inexpensive to Measure	May not reflect individual performance
	Intensive Care Unit Staffing	Strong predictor of outcomes for rare, complex operations	Difficult to make actionable
Process			
	Administration of preoperative antibiotic	Straight forward to measure and track	Research still needed to find the best process to target
	Removal of foley by post-operative day 2	Readily Actionable	Adherence does not always translate to better outcomes
Outcomes			
	30 day mortality rate	Strong face-validity with surgeons	Requires large sample sizes to detect meaningful differences
	surgical site infection rates	Reflect the "bottom-line" as seen by most stakeholders	Expensive to collect data for accurate measures

Structure

Structure refers to the measurable elements of a hospital or provider. Examples include hospital size (e.g. number of beds), provider characteristics (e.g. years in training, annual operative volume), or resource availability (e.g. presence of an intensive care unit.) The most attractive aspect of using structure as a quality measure is that the data is highly objective and easily collected. The structure approach to quality was used to describe that higher volume hospitals have improved outcomes for pancreatic resection. While this approach does give important information to compare hospitals broadly, it provides little actionable information for individual improvement within a hospital.

Although the structure approach is helpful to policy makers or payers comparing broadly across multiple hospitals, surgeons pursuing quality improvement within their department or division may find little use for this approach. This approach to measurement does not identify specific pathways for improvement, other than redesigning structures to meet the quality benchmarks. Needless to say, changes to annual volume, hospital beds, or other resources are difficult to make. Many argue that we need to understand why these structural measures are associated with better quality and "export" these details to other facilities. However, an inventory of practices that distinguish high volume from low volume providers, for example, has not been identified.

Process

Process describes the details of care that can be measured. Examples include giving preoperative heparin for thromboembolism prophylaxis or removing an indwelling bladder catheter by post-operative day two to prevent urinary tract infections. Use of process measures has many practical advantages, particularly for local quality improvement initiatives (Bilimoria 2015). First, process measures are straightforward to collect and track. They typically involve a binary variable (e.g. preoperative antibiotics given – yes or no) that does not require risk-adjustment. Second, process measures are highly actionable. While an outlier outcome (e.g. increased surgical site infection rates) may signal a need for quality improvement, a review of adherence to process measures (e.g. preoperative antibiotics, skin decontamination, appropriate bowel prep) directly identifies steps in care that could be improved.

Although process measures are easier to measure and readily actionable, they do have limitations. Because surgical outcomes are often multifactorial, adherence to a process measure does not guarantee an improvement in outcomes. Additionally, because surgical care is complex and so many processes are involved, it can be challenging to identify the best process measure to track and target. As more research develops to find the "right" process measures that best improve outcomes, it will be easier to use this approach to facilitate quality improvement. For departments or hospitals in the early stages of quality improvement, process measures provide an effective and relatively low resource burden approach to begin evaluating and improving performance.

Outcomes

Outcomes represent the end result of care provided. Examples include rates of mortality, complications, or reoperation. A major benefit of this approach is that it enjoys a high degree of clinical-face validity with most surgeons. Unlike approaches to process and structure, however, outcomes are more difficult to measure and therefore require more resources. For example, when comparing process measures between providers one would simply need to determine the, "% of patients who appropriately received perioperative antibiotics." To compare the outcome (i.e. surgical site infection), however, one would need to collect the rate of the event as well as data about the patient specific risk factors for that outcome (e.g. history of surgical site infection, steroid use, type of operation) to allow for fair risk adjustment. Additionally, even with that added information, sufficient volume is needed to avoid Type 1 and Type 2 statistical errors (discussed in detail below.)

While measuring outcomes may be the most challenging and resource intensive, it has been widely adopted by surgical societies, divisions and departments. This likely reflects that surgeons identify with this form of measurement as relevant to practice. Many are optimistic that the burden of collecting data from electronic medical records will become easier and lead to more efficient risk-adjustment models. Even with improved efficiency, however, most outcomes measurement programs in surgery are still constrained by limited sample size (particularly in local quality improvement efforts) to make them useful in detecting differences in quality, as discussed in detail in the next section.

Understanding Statistical Constraints When Measuring Outcomes

Given the complexity of measuring quality using surgical outcomes, and the large number of outcomes measurement programs in surgery, we will devote this section to understanding the statistical nuances of studying variations in surgical outcomes.

All outcome measures will display some variation. Most often, the variation is attributed to the quality of care provided by the hospital or surgeon. There are, however, multiple other reasons for variation to occur related to chance and case-mix that should also be considered. We review them in a surgical context here.

The Role of Chance

When evaluating surgical outcomes, chance can lead to important flaws in inference, known as Type 1 and Type 2 errors. Both of these errors occur more often with low volume procedures (e.g. pancreatic resection) or when with adverse events are

rare (e.g. death after a cholecystectomy.) Because many of the procedures performed by surgeons are relatively rare, or have infrequent adverse events, understanding the role of chance is essential when assessing outcomes.

Type 1 Errors

A Type 1 error occurs when outliers– *good or bad* – are due to chance. Consider, for example, two surgeons who perform pancreatic surgery. While a death after pancreatic resection is relatively rare, if the death occurs in a surgeons first five operations, he or she will inaccurately be labeled with a "20 %" mortality rate. Similarly, a surgeon who performs five pancreatic resections without a death will also be misleadingly labeled with a "0 %" mortality rate. The difference in mortality rates observed between those two surgeons is more likely due to chance in the setting of low sample size, rather than either exemplary or substandard care.

The so-called "zero-mortality paradox" observed in a study using Medicare claims data provides a useful example of a Type 1 error observed at the hospital level (Dimick and Welch 2008). Researchers reviewed patients undergoing pancreatic resections and identified hospitals with a "0 % mortality" rate for 3 years. Paradoxically, the following year, those hospitals had mortality rates 30 % higher than other hospitals. How could that be? On further evaluation it became clear that those hospitals simply had low volume and "good luck" that led to them inaccurately being labeled high quality. They simply had not performed enough cases yet to have a bad outcome.

Type 2 Errors

A Type 2 error occurs when differences in quality are not detectable due to limited sample size. Although Type 2 errors are widely recognized in clinical trials (e.g. the study was "underpowered"), they are commonly overlooked in surgical quality improvement efforts. Consider, again, two surgeons who perform pancreatic resections. While there may be real quality differences in the care they provide, it would be difficult to identify after only 10 operations. After 100 operations, it would become more clear and with 1000 operations even more so. The ability of increased sample size to help detect differences between two groups is often referred to "statistical power."

Consider, for example, the *Agency for Healthcare Research and Quality* initiative to use post-operative surgical mortality rates of seven complex operations – coronary artery bypass graft surgery (CABG), repair of abdominal aortic aneurysm, pancreatic resection, esophageal resection, pediatric heart surgery, craniotomy, hip replacement – to identify differences in hospital quality. While principally this made sense (i.e. hospitals with higher mortality rates for a given procedure are likely providing lower quality of care), this determination can only be made if there

is enough surgical volume (i.e. enough statistical power) to identify differences. For example, to detect a doubling of mortality among hospitals performing esophageal resections, each would need to complete at least 77 per year. When researchers evaluated all seven of these procedures in the *Nationwide Inpatient Sample*, they found that for 6 of the 7 procedures (all except CABG) the vast majority of hospitals did not perform a minimum case load to detect a doubling of mortality rates (Dimick et al. 2004). In other words, even if there were real differences in quality between hospitals for those operations, low volume would prevent them from being detected.

The Role of Case Mix

Many surgeons confronted with a report identifying them as a poor quality "outlier" argues that their patients are sicker, i.e., that their case-mix is different. Case-mix refers to the type of patients and the type operations being performed. Without question, surgeons and hospitals taking care of sicker patients and doing more complex procedures have a more challenging case-mix that should be acknowledged when we measure outcomes.

The role of case-mix influencing observed outcome rates is most apparent when comparing groups with significant underlying differences. For example, if we wanted to evaluate the mortality rates at a small community hospital performing elective outpatient surgery versus a large tertiary academic center that takes on complex inpatient operations case-mix becomes very important. Even if the same quality of care is provided in both settings, we would still expect a contrast in mortality rates due to differences in patient severity of disease and complexity of the operations performed-- differences in case-mix.

In contrast, when comparing patient populations that are relatively homogenous undergoing similar procedures, accounting for case-mix has less impact. Consider, for example, the mortality rates in the state of New York for patients undergoing coronary artery bypass grafting. The unadjusted rates varied from <1 to 4 %. After adjustment for case-mix, the variation remained the same and was highly correlated (Dimick and Birkmeyer 2008). This should not be surprising because the patients were undergoing the same procedure, had similar underlying diagnoses, and by nature of the disease relatively similar age and health demographics. Thus, the more similar comparison groups are, the less important case-mix becomes in accounting for variation.

For most internal quality improvement efforts, the influence of case-mix on outcomes will be small. Most surgeons' practices' and hospitals systems have little variability year to year in the type of procedures performed or severity of patients being served. While acknowledging variation in case-mix can help establish buy-in from other peer surgeons, minimal resources should be devoted to complex risk-adjustment strategies for local quality improvement projects, unless significant shifts in case-mix are suspected.

Emerging Measurements of Quality in Surgery

While structure, process and outcomes will remain prominent approaches to measuring quality, other measurement strategies are likely to become more visible in the future including patient reported outcomes and surgical video (Table 3.2).

Patient Reported Outcomes

Patient reported outcomes (PROs) include information obtained directly from a patient about their health experience. Examples include patient self-administered questionnaires or focused interviews. They can serve multiple purposes toward measuring and improving the quality of surgical care. First, PROs can further describe the impact of surgery on the patient by soliciting information not captured in our traditional outcomes (e.g. asking about activity levels after a large hernia repair.) Second, PROs may help providers detect when additional interventions are needed (e.g. a patient reported low mobility score prompting a physical therapy evaluation.) Third, PROs may be reported back to individual providers to identify potential patterns of care that could be improved (e.g. provider with consistently high pain scores may re-evaluate his or her post-operative pain regimen.)

Although it is intuitive that we should solicit and incorporate the patient's perspective into how we measure quality and improve care, how to do it well and fairly remains a challenge (Bilimoria et al. 2014). Methodologic issues to be addressed include standardizing questions, integrating the information into already existing medical records and identifying a source of funding for data collection. If PROs are used by payers and regulators to assess providers, then additional research will also be needed to develop appropriate "risk adjustments" for differences in case-mix.

Table 3.2 Emerging measures of quality in surgery: patient reported outcomes, surgical video

Type of measure	Example	Benefits	Drawbacks
Patient Reported Outcomes			
	Generic or disease specific quality of life instruments	Understand outcomes from patient perspective	Instruments and methodology largely unexplored
		Identify gaps in care from a patient perspective	Very burdensome data collection
Surgical Video			
	Video ratings based on skill and technique	Focuses on the quality of the operation, which is understudied	Resource intensive to collect, edit and review surgical video
	Video based peer coaching	Can provide surgeon specific feedback for improvement	Evidence limited to a few procedures (e.g., bariatric surgery)

While methods to obtain PROs are being developed and refined, they will likely have the most uptake initially for operations that our traditional quality measures do not stratify well because they are low volume (e.g. hand surgery) or have low adverse outcome rates (e.g. inguinal hernia repair.)

Use of Surgical Video

While all the previous discussed measurement approaches evaluate what happens to the patient before or after an operation, surgical video uniquely focuses on the operation itself. Because most of our surgical platforms (e.g. laparoscopy, endoscopy) are now fitted with built-in video recording technology, visual data is readily available to surgeons. Early reports in bariatric surgery have been able to correlate an individual surgeon's objective technical skill scores during an operation to his or her patient's post-operative outcomes (Birkmeyer et al. 2013). In doing so, a tremendous amount of interest has been generated in how video data can be used to measure and improve surgical quality.

Multiple possibilities exist for surgical video to be integrated into how we measure and improve surgical quality. For example, individual surgeons are now participating in coaching trials where they watch their own surgical video with a trained peer (i.e. "a coach") to identify where technique can be improved (Greenberg et al. 2015; Hu et al. 2012). In doing so, an entire new range of variables are being identified (e.g. handling to tissue, type of stapler, efficiency of sewing) that may become important measures of quality. In addition to the potential for improving individual surgical quality, if video observed measures are consistently linked to patient outcomes, they may be readily incorporated into surgeon accreditation and board certification. At present, use of surgical video is resource intensive and has only studied for a limited number of procedures.

Choosing the Right Approach to Measure Quality

Recognizing that there are limited resources for quality improvement, it can be difficult to choose where efforts should be prioritized. While there are judgement calls and local limitations about which approach – structure, process, outcomes – can be implemented to measure quality, there are also very real statistical limitations that need to be considered.

Choosing the best measurement approach for quality should take into account the nature of the procedure and our ability to detect differences in what we measure. From a statistical perspective, the more often an event occurs, the easier it is to detect. Therefore, to measure quality for a given procedure we need to ask:

1. How often is the procedure performed? (i.e. Is it high or low volume?)
2. How often does the adverse event occur? (i.e. Is it high or low risk?)

These two questions can guide us to choosing the right approach (Fig. 3.1). Consider the following four categories based on an operations volume and risk.

High Volume, High Risk Procedures

Operations that are high volume and high risk should be evaluated using an outcomes measurement approach. Common examples in this category include bariatric surgery, cardiac surgery, or colectomy. Since these operations are performed commonly and also have relatively frequent adverse outcomes (e.g. colorectal operations surgical site infection rates as high as 30 %) there is enough statistical power to detect differences in *outcomes*.

High Volume, Low Risk Procedures

Process measures and patient reported outcomes are best utilized for procedures that are high volume but have low risk. A common example is the inguinal hernia repair. Although it is one the most common general surgery procedures performed, complications in outcomes are rare (e.g. 30 day mortality <1 %). Differences in patient reported outcomes (e.g. post-operative pain, quality of life measures) or adherence to process measures (e.g. appropriate use of perioperative beta-blocker medication) occur more frequently and therefore more useful when looking for differences in quality.

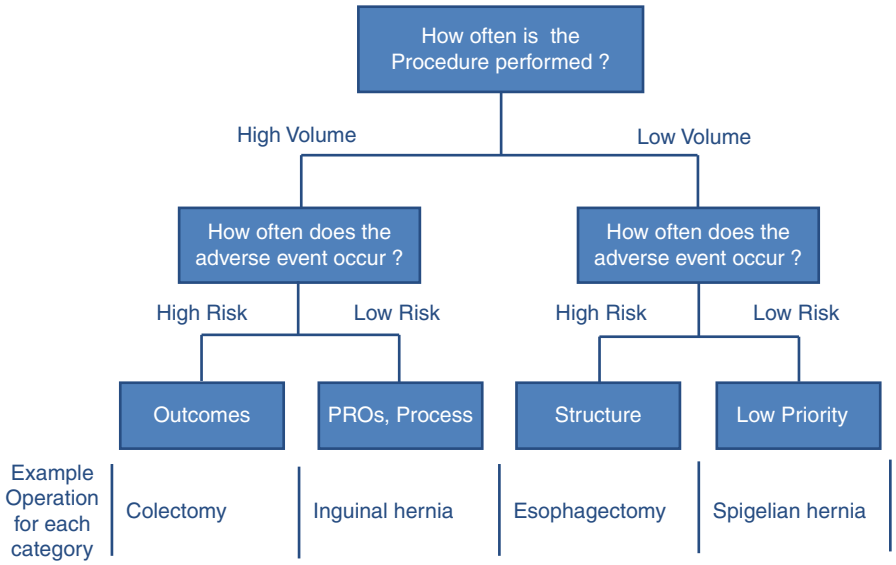

Fig. 3.1 Choosing between different approaches to measuring surgical quality

Low Volume, High Risk Procedures

When procedures are low volume and high risk, structure is the most appropriate approach to evaluating quality. Esophageal resection is an illustrative example. Although it is an operation with a high complication rate, it is performed relatively rarely such that we do not have enough statistical power to detect meaningful differences in either outcomes or process measures. To date, our best evidence now supports structure elements (e.g. operative volume) as the best empiric predictor of quality and future performance (Birkmeyer et al. 2006).

Low Volume, Low Risk Procedures

Finally there are procedures that fit neither outcomes, process or structure approaches because they are low volume and low risk. Operations in this category (e.g. Spigelian hernia) should be given low priority for quality improvement initiatives in favor of operations mentioned above that are more common or have more risk.

Additional Considerations for Measuring Outcomes

Measuring outcomes has the most face-validity with surgeons interested in quality improvement, but is also the most challenging to do fairly. As described above, variation in observed outcomes may be highly influenced by case-mix and chance that can be partially accounted for with different statistical techniques including risk-adjustment and reliability-adjustment.

Risk Adjustment

Risk adjustment can help account for variation that is due to differences in case-mix. This is most frequently done with a multivariable logistic regression model that uses measurable differences in patients (e.g. age, gender, race) to adjust their risk for an outcome. How much "adjustment" is needed depends on the underlying differences between the groups being compared. Although there is eagerness to include as many as 21 "adjustment variables", for procedure specific comparisons (i.e. comparing patients undergoing the same operation) as few as 5 variables can be used to provide the same amount of adjustment (Dimick et al. 2010). This is important to identify because the differences in resources required to collect 5 versus 21 data points about each patient can be significant burden. Again, for internal quality improvement efforts where there is low variability in case-mix, risk-adjustment should not be prioritized.

Reliability Adjustment

Reliability adjustment can help evaluate if the variation observed is due to true differences in performance or instead "statistical noise" caused by low sample size. This approach is more complicated, but briefly uses hierarchical modeling and Bayesian methods to average an individual surgeon outcome rate with the outcomes rate of all the surgeons combined in a weighted fashion based on volume. In this model, a surgeon who has performed zero operations is assumed to be whatever is the average rate of all the surgeon in that group until he or she performed enough operations to stratify themselves are either a high or low performer. The weighted nature of this model results in lower volume surgeons being "adjusted" closer to the mean. While this reliability adjustment can prevent inaccurate labelling of low volume providers as high or low outliers, it also results in a "shrinkage" in the observed variation making it more difficult to detect differences in quality that exist.

Conclusion

Understanding how to measure quality in surgery is necessary for performance improvement. The role of chance when sample sizes are small is often overlooked and can mislabel surgeons inaccurately into high and low providers. Adjustments for case-mix, although available, are resource intensive and for many local quality improvement efforts not necessary. If case-mix adjustments are applied, every effort should be taken to stream line them to a limited number of variables. In addition to our current measures of process, structure and outcomes, new sources of data such patient reported outcomes and surgical video will likely be integrated into the mainstream of quality assessment. Finally, and most importantly, measurement of quality is necessary, but not sufficient for quality improvement. All efforts to measure performance should be coupled focused interventions to improve care.

References

Bilimoria KY. Facilitating quality improvement: pushing the pendulum back toward process measures. JAMA. 2015;314(13):1333–4.

Bilimoria KY, Cella D, Butt Z. Current challenges in using patient-reported outcomes for surgical care and performance measurement: everybody wants to hear from the patient, but are we ready to listen? JAMA Surg. 2014;149(6):505–6.

Birkmeyer JD, Dimick JB, Staiger DO. Operative mortality and procedure volume as predictors of subsequent hospital performance. Ann Surg. 2006;243(3):411–7.

Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369(15):1434–42.

Dimick JB, Birkmeyer JD. Ranking hospitals on surgical quality: does risk-adjustment always matter? J Am Coll Surg. 2008;207(3):347–51.

Dimick JB, Welch HG. The zero mortality paradox in surgery. J Am Coll Surg. 2008; 206(1):13–6.
Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. JAMA. 2004;292(7):847–51.
Dimick JB, Osborne NH, Hall BL, Ko CY, Birkmeyer JD. Risk adjustment for comparing hospital quality with surgery: how many variables are needed? J Am Coll Surg. 2010;210(4):503–8.
Donabedian A. The quality of care. How can it be assessed? JAMA. 1988;260(12):1743–8.
Greenberg CC, Ghousseini HN, Pavuluri Quamme SR, Beasley HL, Wiegmann DA. Surgical coaching for individual performance improvement. Ann Surg. 2015;261(1):32–4.
Hu YY, Peyre SE, Arriaga AF, et al. Postgame analysis: using video-based coaching for continuous professional development. J Am Coll Surg. 2012;214(1):115–24.