Quality and Outcome Assessment for Surgery

Laurence Chiche, MD, Han-Kwang Yang, MD, PhD,† Fariba Abbassi, MD,‡ Ricardo Robles-Campos, MD,§ Steven C. Stain, MD,‖ Clifford Y. Ko, MD, MS, MSHS,¶ Leigh A. Neumayer, MD, MS, MBA,# Timothy M. Pawlik, MD,** Jeffrey S. Barkun, MD,†† and Pierre-Alain Clavien, MD, PhD‡✉*

Abstract: This forum summarizes the proceedings of the joint European Surgical Association (ESA)/American Surgical Association (ASA) symposium on Quality and Outcome Assessment for Surgery that took place in Bordeaux, France, as part of the celebrations of the 30th anniversary of the ESA. Three presentations focused on a) the main messages from the Outcome4Medicine Consensus Conference, which took place in Zurich, Switzerland, in June 2022, b) the patient perspective, and c) benchmarking were hold by ESA members and discussed by ASA members in a symposium attended by members of both associations.

Keywords: quality, outcome assessment, consensus conference, perspective of patients, benchmarking

(*Ann Surg* 2023;278:647–654)

As the universal outcome assessment of surgical procedures remains a challenge worldwide, the European Surgical Association (ESA) and American Surgical Association (ASA) took the task to highlight the main statements from a Jury-based consensus conference on how to assess the quality of surgical interventions. Thereby special emphasis was provided on the perspective of patients, the most important stakeholder, and on appropriate benchmarking for comparisons among centers or various therapies. What follows is a compendium of presentations hold by leaders in the field of perioperative outcome research from both of our continents followed by robust discussions that took place at the May 11 meeting in Bordeaux, France.

OUTCOME ASSESSMENT FOR SURGERY

Pierre-Alain Clavien (Zurich, Switzerland)

Grounding Question

Statements from the Outcome4Medicine Consensus Conference, which took place in Zurich, Switzerland, in June 2022.[1]

In 2008, the World Health Organization (WHO) recognized that postoperative complications account for a large proportion of preventable medical injuries and deaths and deemed them a global public health issue.[2] The lack of standardized and universal surgical endpoints has led to inconsistent, arbitrary, and often clinically irrelevant outcome assessments, opening the door to biased interpretations, also hindering the improvement of health care quality.[3,4]

To address this issue the Outcome4Medicine core group engaged in a long process aimed to produce consensus statements on how to assess the outcome of medical interventions. For this purpose, *the Zurich-Danish Model* was chosen, where an independent Jury frames recommendations based on the best available evidence prepared ahead of the meeting by a multidisciplinary panel of experts.[5] The aim was to produce evidence-based, internationally valid, and unbiased recommendations, which consider the perspectives of many stakeholders, including patients, health care providers, as well as payers or governments. To prevent any conflict of interest, Jury members were not directly involved in surgical or medical outcomes research.

The Jury recommendations are summarized in Table 1, as published in the original *Nature Medicine* paper.[1] First, the Jury suggested standardized time points for recording outcome assessments and moving away from historically collected discharge or 30-day data only.[6,7] Five fixed time points were proposed, ranging from the predisease state (T0) through perioperative state (T1–T3) to the long-term follow-up at 5 years postoperatively (T4).

The next focus was on assessing the postoperative course from the perspective of the health care providers with much emphasis on morbidity rather than mortality. Consistent and complete reporting on complications is paramount including the availability of an objective grading system ranking complications by severity. The Jury recommended a listing and grading of complications according to severity using the Clavien-Dindo classification,[8,9] and to capture the cumulative morbidity for a single patient with the Comprehensive Complication Index (CCI®), ranging from the value 0 (no complication) to 100 (death of the patient).[10,11] Both approaches have already been well-established and validated in many fields of surgery. Their consistent use in the clinic and in publications would enable standardization of reporting morbidity,[12,13] and the Jury deemed the use of these 2 tools as a minimum when assessing postoperative outcomes. The so-called "failure to rescue" is another important marker of quality indicating a delay or failure in recognizing and properly managing postoperative complications often with a fatal outcome.[14–16] This highly relevant marker of quality often favors large-volume centers with the consistent availability of interdisciplinary experts.

Points 3 and 4 of the Jury recommendations deal with patient's perspective and benchmarking, 2 topics that Prof. Laurence Chiche and Prof. Han-Kwang Yang will discuss in more details, respectively, in the second and third part of this jubilee article.

From the *University Hospital of Bordeaux, Bordeaux, France; †Seoul National University College of Medicine, Seoul, Republic of Korea; ‡University Hospital of Zurich, Zurich, Switzerland; §Virgen de la Arrixaca Clinic and University Hospital IMIB, Murcia, Spain; ‖Lahey Hospital and Medical Center, Burlington, MA; ¶University of California at Los Angeles and the American College of Surgeons, Los Angeles, CA; #University of Florida, College of Medicine-Jacksonville, Jacksonville, FL; **The Ohio State University, Wexner Medical Center, Columbus, OH; and ††Department of Surgery, Faculty of Medicine, McGill University, Montreal, QC, Canada.

✉clavien@access.uzh.ch.

The authors report no conflicts of interest.

ISSN: 0003-4932/23/27805-0647
DOI: 10.1097/SLA.0000000000006077

TABLE 1. Jury Recommendations From the Outcome4Medicine Consensus Conference

Timepoint of outcome assessment
T0: Predisease state
T1: Before intervention
T2: Early postoperative phase (short-term)
T3: Mid-term follow-up (intervention and disease specific)
T4: Five-year follow-up
Outcome assessment for postoperative complications
When measuring postoperative complications use at a minimum:
Clavien-Dindo classification (severity of complications)
Comprehensive Complication Index (CCI®) (total morbidity)
Failure to rescue
Conduct regular, interdisciplinary "morbidity and mortality conferences" in clinical practice and include discussions of successful cases to reinforce effective behavior.
Patient-centered outcome assessment
Patient education and empowerment are critical for them to take on responsibility for their care.
Incorporate patient-centered outcome measures such as PROMs and PREMs into routine clinical care.
Use, at a minimum, one standard global life satisfaction measurement (eg, EQ-5D) to quantify change over time.
Benchmarking
Benchmarking is a prerequisite to assess and improve quality of care.
Make benchmarking mandatory for all institutions, irrespective of their size.
A robust methodology to create benchmark values includes low-risk patients treated at expert, high-volume centers.
Risk assessment
Make preoperative assessment and postoperative reporting of high-risk patients mandatory and disease specific.
Consider multiple factors such as patient-, physician- and procedure-related factors as well as the context in which patients live (for example, socio-demographic factors, social determinants of health).
Involve "informed" patients in the establishment of risk profiles and in defining their expectations of the intervention.
Data management
Governments and relevant sectors of society must engage in establishing nationwide health care databases, working toward global standardization of data collection, curation, storage, and validation.
Hire/appoint a "data quality guarantor" at every institution to ensure data accuracy and completeness.
Involve patients in database creation and data collection so that outcome data can be shared with patients in a simple, lay-language manner.
Inform the public about the benefits of well-designed research under strict privacy protections.
Call on the WHO and the G20 and their global responsibility to master the political process and make the creation of anonymized centralized data centers a priority.
Approaches to secure adequate treatments
Implement initiatives to avoid over and undertreatment, such as:
"Choosing Wisely" initiatives
Binding guidelines
Removing financial incentives for low-value interventions
Dealing with cultural and demographic factors
Define and incorporate cultural and demographic factors in the interpretation of outcomes after surgical interventions.
Strategies to deal with unwarranted outcomes
Shift from a culture of blame to a culture of collaborative and collective learning.
Develop and establish new systems and procedures to mitigate the consequences of unwarranted outcomes (for example, compensation at institutional, regional, and/or national level).
In the handling of medical errors, apply truthful disclosure and the TRACK principle: Transparency, Respect, Accountability, Continuity, and Kindness.

With permission from *Nature Medicine* paper.[1]

Further points of the Jury recommendations account for cultural and demographic factors when evaluating outcomes, for example, by directly adjusting outcome measures to culture or by including socio-cultural factors into the interpretation of data.[17–19]

The robustness of any outcome data, irrespective of the perspectives, falls with reliable data management. The goal should be to create nationwide health care databases to move toward a global standardization of data collection, storage, and validation. All stakeholders should contribute to this overarching goal, from patients to health professionals to governments. The Jury even recommended calling to the responsibility of the WHO and G20 to master the political process and declare the creation of anonymized central data centers a priority.

Although the focus is on optimizing outcomes, this does not mean that all available resources should be tossed around to achieve the optimum. Rather, as part of shared decision-making with patients, resources should be used wisely and in the best interest of the patient and society.[20] With these recommendations a framework for outcome assessment is now available that incorporates the perspectives of various stakeholders, with a particular emphasis on patients.

ASA Discussant: Clifford Y. Ko (Los Angeles, CA)

Thank you for this great honor to discuss the work of Dr. Clavien and his experts and jury members of the consensus conference. This project that focuses on outcomes for medicine is very important for us in surgery because identifies and defines the data needed to elucidate how the evaluation of high-quality care might be undertaken. As Dr. Stain the moderator shared, I have been Director of the quality programs at the American College of Surgeons for the last 15 years. When we think about what we're trying to achieve in surgical quality in the United States and Canada (and internationally) we frame our efforts as 3 components: problems, opportunities, and what "good" looks like. So, I'm going to take this opportunity to ask Dr. Clavien about 3 problems that we currently have.

First, surgery has always been focused on complications. However, there are many domains of quality, 2 of which are safety as well as patient-centeredness with patient-reported outcomes. However, in the United States and some others, we know increasingly more that the timeliness of care, where there are waiting lists; efficiency of care (eg, how much money does it take to do a workup, are we spending too many CT scans, etc.); effectiveness (eg, if we do surgery, are we sufficiently curing the clinical problem?); and equity (eg, are treatments and outcomes equitable). These domains are all quality-related issues that need to be measured. So, my first question is: now that Outcome for Medicine is completed, do we need an Outcome for Medicine Part 2 to define perhaps more domains of quality?

Second, in the United States, we also have the problem of implementation of the data. How do we measure the data well in each of our 5000 hospitals? We have a program called

the National Surgical Quality Improvement Programs, or NSQIP. Over decades, we found having accurate data is paramount to achieving high-quality care and outcomes. But right now, collecting accurate data is sometimes expensive and burdensome – and not altogether comprehensive. We have hundreds of hospitals in the United States, but not all. So, my second question is: how to operationalize and implement the use of data that have been identified by the Outcomes for Medicine project everywhere – not just in one country, but all such that benchmarking, evaluation, and improvement can be better undertaken? Will it be through technology and AI?

Third, something we are working on right now is achieving improvement. We know that when audit and feedback is implemented, it works only 16% of the time. This has been shown through many systematic reviews. This is not sufficient. We need to improve improvement. So, my question is: when we provide data and with benchmarks, how do we help hospitals and surgeons get better? What should our strategies be to achieve improvement when achieving improvement is so difficult?

Response From Pierre-Alain Clavien (Zurich, Switzerland)

Thank you so much, Dr. Ko, for accepting our invitation and for providing such great insights into the complex and controversial topic of outcome research. In fact, you already answered most of your questions. Your first question is whether we need an Outcome4Medicine, Part 2, to define more domains of quality. I can only agree with this proposal, as quality must cover other processes, such as the time it takes to see an expert, to have the workup completed, or to be fully informed about the putative therapies. The path to diagnostic testing is also rarely standardized, and it is not typically well-described in the guidelines since the final diagnosis remains unknown. This could be a valuable consensus effort, combining both ASA and ESA perspectives.

Your second question targets the challenges of obtaining comprehensive and conclusive data from all hospitals in a country. Some countries, such as in Scandinavia, have done better than others, but, in many regions, only an administrative database is available, which lacks key medical information. Despite much effort, I can see that there is still dissatisfaction in the United States. In Outcome4Medicine, the jury recommend independent, and if possible, national data collection. Government must invest in this issue since no improvement in health care can occur without such information. Will AI help? This, I do not know.

Finally, improvement is, of course, the goal. Probably, the main benefit is for us to simply be aware of our results and the gap in benchmark values. There are convincing studies which demonstrate that spectacular self-improvements occur simply when feedback is provided to hospitals and surgeons. It seems helpful to convincingly inform surgeons where they rank compared with their peers. In addition, providing a concrete action plan, illustrating the steps to be taken, to the surgical team has shown promise, although concrete actions for improvements remain in the hand of the respective hospital and doctors. Another effective tool for the improvement of surgical procedures is video-assisted feedback, perhaps with the help of AI. All methods can only be activated once reliable data are available.

PATIENT PERSPECTIVE

Laurence Chiche (Bordeaux, France)

Grounding Questions

Which patient reported outcome should be used? What is the importance of PROMs and PREMs?

The most important stakeholder in outcome assessment are the patients, which are regrettably often forgotten. The mission of every physician is to provide patients with the best care. However, patients define "the best" and quality by other criteria than surgeons, for example, by their functional status postoperatively or by nonmedical services such as food quality. Therefore, to assess the quality of surgery in research and in clinical practice, standardized assessment of outcomes from the patient's perspective including patient-reported outcome or experience measures (PROMs and PREMs, respectively) are needed and should become an integral part of any surgical outcome assessment.[21–23] PROMs and PREMs allow evaluating the real impact of surgery on a patient's life and people's satisfaction[23] and must be validated, according to a rigorous methodology.[24]

PROMs are questionnaires completed by patients, without professional interference, using paper or electronic applications. They are either generic with questions related to pain, general quality of life, social impact, or condition-specific including questions focused on the disease's symptoms or consequences of the surgical procedure.[25] PREMs are surveys completed by patients to assess their experiences. They include questions related to communication with health care providers, timeliness of care, and overall satisfaction with the care provided.[26]

Since the 70s, many different PROMs and PREMs have been formalized by academic, governmental public health institutes, research projects, or user associations all over the world for different fields of medicine.[27–29]

The use of PROMs and PREMs in surgery is increasing for several reasons:[30] first, surgical morbidity and mortality have decreased while long-term survival has improved, leading to focus on quality of life; second, hospital stays tend to be shorter (enhanced recovery and outpatient programs),[31] with more and more home cares; third, alternative and competitive treatments to surgery are becoming more and more popular, making it necessary to compare and better evaluate the different options.

PROMs and PREMs were initially used in low morbidity, functional surgery, and recently because of multimodal associated treatments, after complex surgery like transplantation or oncological procedures.[32–35]

The choice of PROMs to be used in research depends on the objective on the study. Generic and condition-specific PROMs are often associated. For example, SF-36 assesses 8 domains, including physical functioning, role limitations, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health.

The actual use and impact of PROMs and PREMs depends on the country. They are commonly implemented in clinical research for comparison and benchmarking, but also, in practice, for certification of institutions, public diffusion, and reimbursement (health care system).[29] The importance of PROMs and PREMs has been widely demonstrated providing valuable information for surgeons, health care providers, and policymakers to evaluate the effectiveness of surgical interventions. By using these measures, surgeons can assess the benefit of the surgery in terms of quality of life and social consequences, better understand postoperative symptoms and their severity.[23,36] So, surgeons can identify areas for improvement in patient care or even in

technique, propose a potential alternative, better inform before surgery and, detect earlier, after the patient's discharge, the late potential side effects to react faster. The benefit for the patient has also been shown: use of PROMs improves symptoms (linked to anxiety) and even survival in palliative management.[37] It also helps to decrease the use of emergency services and the rate of readmission. Concerning the Health system organization, it helps to identify areas for improvement in institutions and the organization of care by implementation of clinical pathways for example, and finally, PROMs happen to be an essential tool to set up medical telesurveillance.

Nevertheless, even if surgeons seem convinced of the benefits of PROMs,[38,39] some barriers to their use, mainly in clinical practice, can be highlighted[40]: (1) patient's acceptability and fidelity in case of too long or irrelevant questionnaires, (2) absence of a definition of acceptable response rate, (3) equity issues because of language problems or disadvantaged backgrounds, (4) cost-effective and time-consuming implementation, (5) lack of standardization of condition-specific PROMs for comparison, (6) actual impact in care modifications.

To conclude, PROMs and PREMs are effective tools to improve the quality in surgery, in the era of minimal invasive surgery, postoperative home care, and telesurveillance. To integrate them into clinical practice is a strong recommendation, as long as they are used to complement clinical data, are validated, reliable, easy to understand, available in different languages, and wisely chosen, considering the targeted objectives and improving health equity.[41]

Of further importance for the Outcome4Medicine Consensus Conference Jury is that patients should move away from their role as background artist and take an active role in choosing their care. Patient education and empowerment are key elements to support patients in taking responsibility for their care.[36,42]

ASA Discussant: Leigh A. Neumayer (Jacksonville, FL)

I would like to thank the ESA and ASA for the privilege of being the first discussant. It's a pleasure to be here. Over the last century, advances in medicine and surgery have increased survival rates for most diseases. Importantly, we have decreased postoperative mortality to low single digits, if not zero, for nearly all elective procedures. Additional surgeon-centered outcomes, such as complications and the ability to rescue patients, can also be measured, although defining these outcomes is not as black and white as mortality. In fact, in the NSQIP, when we were doing this work on data definitions, both for the comorbidities and complications, it was hard. We started to define them based on what kind of intervention was done because that's easier to measure. Death is easy to define, but the cause of it is more difficult. From the patient's perspective, as we saw from some of the examples already given, their quality of life after the procedure and their satisfaction with the process are what matters. Interestingly, we've fully embraced goals of care discussions for patients at the end of life. Yet, we've not incorporated them into discussions with patients when deciding on whether to operate or which procedure may suit their goals best.

The importance of using both surgeon-centered measures and patient-centered outcomes was shown in the Veterans Affairs trial of open or laparoscopic inguinal hernia repair.[43,44] In this trial, we randomized 1,983 men to open or laparoscopic mesh repair of hernia, with the primary outcome being surgeon-centered, which was recurrence at 2 years. The trial included several patient-reported outcomes including Medical Outcomes Study Short Form 36 (SF-36), surgical pain score (SPS), activities assessment scale (AAS), and patient satisfaction. Two of these measures (SPS, AAS) were developed and validated to specifically assess pain and activity levels after inguinal hernia repair for this study.[45,46] While we found that the recurrence rate in the laparoscopic group was almost 2 times as high as in the open group, we also wanted to better understand what a recurrence meant to the patient versus other complications.[47] What we found was that neuralgia, or pain, adversely affected all patient outcomes, while recurrence, which we said was the gold standard, affected pain activity and satisfaction, but interestingly, not their score on the SF-36. This led us to conclude that we should not only use recurrence as a measure of the effectiveness of hernia repair but also the occurrence of postoperative pain.

Dr. Chiche noted a lot of barriers to us implementing PROMs and PREMs, including patient acceptability and fidelity; the absence of a definition of an acceptable response rate, which we've not really addressed; equity issues due to different languages; cost-effectiveness; the time-consuming nature of implementation; lack of standardization; condition-specific PROMs for comparison; and actual impact of care modifications. In the hernia study, we showed how important measuring PROMs is.

I have 3 questions for Dr. Chiche: First, how important are these condition- and procedure-specific outcome measures, when we have some good data that show that more generic tools like the SF-36 might pick up what the problems are?

Second, I believe that if we're going to use PROMs to change practice, then we're going to need surgeons to believe that they are important. Do you have any data that would show us that, besides our own anecdotal experiences?

Finally, in the United States, the way to get something incorporated into practice is to get it reimbursed by insurance. Any ideas of how to accomplish this?

Response From Laurence Chiche (Bordeaux, France)

Thank you for your interesting comments and these very relevant questions, even if I do not have all the answers. Concerning the condition-specific outcome measures, I think they are much more informative than the generic ones because they focus on the real side effects of the specific procedure, and they could help to improve our technique and the information we deliver before surgery. Ideally, both generic and condition-specific measures should be used.

Regarding the perception of PROMs in the surgical community, you are right. It is paramount that we convince surgeons that they are useful. When preparing this talk, I asked many of my colleagues what they thought of PROMs, and they didn't exactly know what they were. In the literature, there is very few data about this issue. PROMs are quite known in functional and outpatient surgery, but much less in oncological surgery. A very interesting paper on this topic has recently been published by Mou et al[38] in this Journal. So, it is our mission to inform and train young surgeons on how to use PROMs and improve their practice.

Finally, concerning the implementation of PROMs, we need to find a way of making them easier to use or even almost compulsory. This is, perhaps, the most difficult aspect. In France, the insurance reimbursement system is different, and this solution is not as applicable as it is in the United States, for example. Perhaps, it could be incorporated in the ranking system of institutions, although it could be dangerous. Rather, I think it should be incorporated at an institutional level, with regular assessments and financial benefits for departments in terms of research staff and technicians. However, of course, all these measures should be discussed according to the health care system of each country.

ASA Discussant: Jeffrey S. Barkun (Montreal, Canada)

I want to thank Dr. Chiche for an excellent review on the topic of PROMs and PREMs in an era where we understand that complications and mortality cannot reflect the totality of the patient perioperative experience.

I would like to come back to the 2 cartoon examples which Dr. Clavien and Dr. Chiche mentioned to us to explain how discrepant 2 different patients' impressions may be, despite similar measured perioperative PROMs. Arguably, I believe this relates to one of our major roles as surgeons: not as a technician, but rather as a manager of patient expectations in the context of a customized patient approach which starts before the operative journey.

PROMs are thought to measure a patient-recorded state at a given point in time. The difference between a baseline PROM and a postoperative PROM (the "perioperative delta PROM") is often the preferred way to measure the change in a perceived perioperative state of health by numerically using the patient as their own control. However, the appreciation of the patient for the perioperative process is arguably just as dependent on their preoperative expectation as it is on the actual magnitude of the "perioperative delta PROM." Two patients may thus display an identical "perioperative delta PROM" yet have very different perceptions of their experience because of their respective expectations. The magnitude of these expectations is, however, never quantified preoperatively.

One way of summarizing numerically a patient's expectation could be to measure the preoperative PROM and ask the patient at the same preoperative session to give us an anticipated postoperative PROM which they would expect to achieve after the operation. The difference between the baseline PROM and the anticipated postoperative PROM could be thought to reflect a measure of the patient's expectation, call it the "expected delta PROM."

After the operation, a simplistic measure of the fulfilled patient expectation could then be a ratio of the "preoperative delta PROM" over the "expected delta PROM" which the patient had hoped for. In an ideal situation, the ratio will show a better-than-expected result.

As surgeons, we daily need to evaluate and respond to patient expectations according to our knowledge of the patient, the literature, the operation, and our experience with previous patients, all in a data-driven fashion. Trying to measure our patients' expectations would be a good place to start.

Response From Laurence Chiche (Bordeaux, France)

Your question is important. I took this example because it's 4 weeks after surgery. If you give the patient PROMs at day 15, you can anticipate the problem. If we use these PROMs routinely, then we can anticipate the problems and better communicate with the patient. We must determine when these PROMs will be used, and it must be easy to respond to them. I believe that this is easier to implement locally than nationally, as it's a heavy procedure. However, this can be done by each institution. At our institution, we built a clinical pathway for this. Importantly, when surgeons realize how patients feel, they can speak with them before the operation. When a patient is prepared, everything is better, including their symptoms and satisfaction.

BENCHMARKING IN SURGERY

Han-Kwang Yang (Seoul, South Korea)

Grounding Question

What is the goal of "Benchmarking" in surgery and which methodology should be used?

Unbiased comparisons of outcomes translate the quality of care among individual therapies, physicians, hospitals, or even health care systems. Novel approaches for benchmarking are central for quality assessment and improvement comparing performances to "the best."[48,49]

Among the outcome measurement from the clinicians' perspective, the Clavien-Dindo classification has been the most widely used classification in the literature. It has been further developed in an index, the Comprehensive Complication Index (CCI®), reflecting the impact of multiple complications (ie, whole morbidity), respecting the degree of severity of each complication. The computation of the CCI® is readily available online for single patients, group of patients, or even institutions; and importantly both the Clavien-Dindo system and CCI® have been selected by the Jury of the recent Outcome4Medicine Consensus Conference[1] and have served as key endpoints in all recent benchmark studies. The prerequisite condition for the proper use of these metrics for comparisons, however, is the accurate data collection. Once available, sharing the outcome among doctors or institutions leads to improvement in the outcome.[50]

In contrast, textbook outcome (TO) index has been proposed to measure how many portions of the operated patients received optimal postoperative course. TO can be defined by multiple parameters such as the absence of intraoperative complication, free tumor resection margin, appropriate of lymph nodes in the resected specimen, no severe postoperative complication, and so on; for example in gastric cancer surgery.[51] It is important to realize that in heterogeneous groups of patients, TO cannot be used for fair comparisons among institutions or countries. In that sense, comparison of outcomes in the best patient set is necessary. Performance of "the best" relates to the fact, that benchmark values for a specific surgical procedure are based on the outcomes of low-risk patients treated in international high-volume reference centers. To create ambitious but achievable benchmarks, the benchmark cutoff is set at the 75th percentile of the center's median. In this way, debates about ambiguous risk adjustment can be avoided. To create a valid benchmark, it is necessary to define the group of patients associated with the lowest risk for complications such as young age, low body mass index, the absence of comorbidities. Eligible centers for benchmark determination should be high-volume centers holding a prospective database, be involved in clinical research in the field of interest, and be from at least 2 continents. A minimal study period of 4 years for benchmarks has been recommended.[48,49] The Jury of the Outcome4Medicine Consensus Conference recommended the implementation of proper benchmarking mandatory for all institutions to enable trustworthy comparisons, as any health care providers must strive to reach "the best" outcomes (Table 1).

While benchmarking originally comes from the fields of economy, it is now widely adopted in medicine.[52,53] Since 2016, such benchmark values have already been determined for > 15 procedures by several groups in various field of general surgery.[52–65]

A possible routine use of the benchmark values is to select patients with outcome parameters outside of the benchmark values for discussion at the institutional multidisciplinary morbidity-mortality conferences. Poorer outcome than benchmark values may relate to deficiencies in care or just because the patient belongs to a higher risk group. A new finding was identified in most benchmark studies in identifying the best centers. In addition to the well-established center-volume or rates of failure to rescue, centers of excellence disclosed a high proportion of higher risk patients (ie, nonbenchmark patients). Thus, the ratio of benchmark/nonbenchmark cases represents a new surrogate marker of quality.

ASA Discussant: Timothy M. Pawlik (Columbus, OH)

I would like to thank the ESA and ASA for the opportunity to discuss this paper. Due to the increased cost of health care delivery and the recognition that outcomes vary among hospitals and practitioners, stakeholders have placed an increased emphasis on obtaining and assessing high-value surgical care. In turn, it is paramount for stakeholders to accurately assess the delivery of quality care across hospital systems. Traditionally, we have used postoperative surgical outcomes, such as mortality, readmission, and length of stay as important measures to assess performance, both at the hospital level and at the surgeon level. However, when individually examined, these metrics fail to fully give a picture of overall quality. To account for these shortcomings, investigators have sought to combine several quality metrics into a single component, as we heard today from Dr. Yang. To this end, his group and others have had increased interest in this so-called TO metric, which is a composite metric that globally represents ideal surgical care.

As a summation of routine collected data, TO allows for a reflection of quality from different domains of care. While TO has the advantage of being customizable to a certain disease process or surgical procedure, the elements included in any definition of optimal or textbook can vary significantly, and sometimes, be quite arbitrary. In turn, for TO to be applied globally, it really needs to be defined in a very specific manner. Another problem with TO is that most factors included in the definition are hospital-based and physician-determined. Most factors are all or none. This is particularly problematic because each element of a TO is probably weighted very differently by patients.

TO also suffers from the lack of the ability to compare performances among providers and centers relative to the ideal standard. To that end, we have heard about benchmarking, which has been recently proposed to define a "best standard" for comparative assessment of high-level performance to drive quality improvement. By referring to a point of reference, the so-called benchmark facilitates comparing ourselves to the very best. Benchmarking is centered around the concept of a continuous cycle of defining the best, comparing the best, and learning from the best. I would argue, today, that, with the data around gastrectomy that Dr. Yang showed us from his hospital in Seoul, your hospital is truly the benchmark for gastrectomy.

Long adopted by other industries, the concept of benchmarking has only recently been applied to medicine. And again, I want to congratulate Dr. Yang for his important presentation and recognize Dr. Clavien for organizing what really was a wonderful Outcomes4Medicine seminar in Zurich last year. While attractive as a concept, benchmarking can be challenging. Benchmarking should involve choosing a well-defined surgical procedure as well as robust, well-defined clinical outcomes. Benchmarking outcomes need to be determined only in the lowest risk patients, and according to strict criteria that are determine a priori. In addition, benchmarking needs to set up numeric standards at those best centers.

I have 4 questions for Dr. Yang: First, what is the best approach to include PROMs and PREMs that we have been hearing about today, in addition to the classic quality metrics that define TO and/or benchmarking?

Second, when determining what factors and what numeric thresholds determine best-in-class or optimal, do we always need a Delphi process, an expert opinion, or a jury? Who and how is "best" defined? We talked a bit about the bias involved in the jury. Perhaps, artificial intelligence will help us define what is "best" in the future?

Third, the elements of any benchmark need to be reconsidered on a routine basis with a broad range of stakeholders. How often do we need a re-benchmark, given the rapid pace of innovation and changes in medicine?

Finally, what are your recommendations around patients, surgeons, and hospital systems, regarding how they should use these metrics, such as TO and benchmarking? Should these data be publicly reported to patients, and should they be used to make decisions around the regionalization of care for certain procedures? Again, thank you for a fantastic presentation on an important topic.

Response From Han-Kwang Yang (Seoul, Republic of Korea)

Thank you very much Dr Pawlik for your comments and questions. Regarding your first question on PROMs and PREMs, I believe that it will be difficult to incorporate them in TO, but it may be easier in our benchmark studies.

Your next question on the "best in class" is difficult to respond to, and I would agree that this can also be somewhat subjective. Probably, expert consensus can be a reasonable way to establish the criteria. However, it should be strictly data-based. In the future, we may use the 4 surrogate markers: quality, center-volume, rates of failure to rescue, and the proportion of nonbenchmark cases. I do not think that a Delphi would help, and perhaps, AI could be useful in the future.

Third, your question regarding how frequently we should repeat a benchmark study is procedure-dependent, possibly related to the availability of other competitive therapies. As I presented, if you consider that the purpose of all these parameters is to provide feedback to the doctors or the institute to improve their results by revealing their performance, then the more frequently you do this, the better it will be. It would be best if these markers are automatically calculated (eg, at our Gastric Cancer Center, we evaluate each surgeon's overall complication rate on a weekly basis, as well as each complication's component value and cumulative values, ie, their respective CCI®).

Regarding your last questions, TO can be more useful in a single institute as a longitudinal comparison because the patient population might not have changed that much within such a timeframe. It is a challenge to grasp how the data gathered through benchmark studies should be used. I am not in favor of automatically releasing information on a public scale; however, I would be in favor of providing this information to each institution for them to observe potential gaps and react to them accordingly. Certainly, benchmarking data comparison is superior to encourage each institution to improve on specific areas. The data can be carefully released to the public anonymously. This information can also be used at the level of the legislator, for example, when working toward the centralization of some complicated procedures.

ASA Discussant: Steven C. Stain (Burlington, MA)

I was struck by the fact that you were able to regionalize the care of gastric cancer patients without mandating it, as it seems patients were aware of which centers had the best outcomes. My question is about the lymph nodes. You can do a gastrectomy routinely and get 30 lymph nodes, but several studies have suggested that <30 lymph nodes would be adequate. Is it proper to increase the threshold to 30 until you get data that 30 has better survival than 15?

Response From Han-Kwang Yang (Seoul, Republic of Korea)

Yes, thank you for asking this important question. When comparing the rate of survival of gastric cancer by stage in the United States versus Korea or Japan, it was found that every stage had a poorer outcome in the United States. Why is this the case? Most likely, it is not because the surgeons didn't perform the proper resection, but, probably, because of a variation in the pathologic assessment, based on 2 aspects leading to understaging. For proper gastric cancer surgery, surgeons should first perform enough lymph node dissections, and second, pathologists should carefully look at as many lymph nodes as possible. As I presented, if you only do Level 1 lymph node dissections for distal gastrectomy, you can get a median of 31 lymph nodes; with Level 2, you can get a median of 13 nodes. To tell whether you have completed a proper radical gastrectomy, a median of 15 lymph nodes is too small. This is an important issue that must be explored in further analyses. This topic can be discussed as an expert consensus topic at the International Gastric Cancer Congress.

REFERENCES

1. Domenghino A, Walbert C, Birrer DL, et al. The Outcome4Medicine Consensus Group. Consensus recommendations on how to assess the quality of surgical interventions. *Nat Med*. 2023;29:811–822.
2. WHO Patient Safety & World Health Organization. WHO guidelines for safe surgery: safe surgery saves lives. WHO Guidelines approved by the Guidelines Review Committee; 2009.
3. Horton R. Surgical research or comic opera: questions, but few answers. *Lancet*. 1996;347:984–985.
4. Clavien PA, Puhan MA. Biased reporting in surgery. *Br J Surg*. 2014;101:591–592.
5. Lesurtel M, Perrier A, Bossuyt PM, et al. An independent jury-based consensus conference model for the development of recommendations in medico-surgical practice. *Surgery*. 2014;155:390–397.
6. Lawson EH, Louie R, Zingmond DS, et al. A comparison of clinical registry versus administrative claims data for reporting of 30-day surgical complications. *Ann Surg*. 2012;256:973–981.
7. Parthasarathy M, Reid V, Pyne L, et al. Are we recording postoperative complications correctly? Comparison of NHS Hospital Episode Statistics with the American College of Surgeons National Surgical Quality Improvement Program. *BMJ Qual Saf*. 2015;24:594–602.
8. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg*. 2004;240:205–213.
9. Clavien PA, Barkun J, de Oliveira ML, et al. The Clavien-Dindo classification of surgical complications: five-year experience. *Ann Surg*. 2009;250:187–196.
10. Slankamenac K, Graf R, Barkun J, et al. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann Surg*. 2013;258:1–7.
11. Clavien PA, Vetter D, Staiger RD, et al. The Comprehensive Complication Index (CCI): added value and clinical perspectives 3 YEARS "Down the Line". *Ann Surg*. 2017;265:1045–1050.
12. Boxhoorn L, van Dijk SM, van Grinsven J, et al. Immediate versus postponed intervention for infected necrotizing pancreatitis. *N Engl J Med*. 2021;385:1372–1381.
13. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med*. 2009;360:491–499.
14. Burke JR, Downey C, Almoudaris AM. Failure to rescue deteriorating patients: a systematic review of root causes and improvement strategies. *J Patient Saf*. 2022;18:e140–e155.
15. Silber JH, Williams SV, Krakauer H, et al. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. *Med Care*. 1992;30:615–629.
16. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. *N Engl J Med*. 2009;361:1368–1375.
17. Hayes S, Napolitano MA, Lent MR, et al. The effect of insurance status on pre- and post-operative bariatric surgery outcomes. *Obes Surg*. 2015;25:191–194.
18. Rohlfing ML, Mays AC, Isom S, et al. Insurance status as a predictor of mortality in patients undergoing head and neck cancer surgery. *Laryngoscope*. 2017;127:2784–2789.
19. Weyh AM, Lunday L, McClure S. Insurance status, an important predictor of oral cancer surgery outcomes. *J Oral Maxillofac Surg*. 2015;73:2049–2056.
20. ABIM Foundation Choosing Wisely Initiative. 2023. https://www.choosingwisely.org/
21. Black N. Patient reported outcome measures could help transform healthcare. *BMJ*. 2013;346:f167.
22. Ko CY, Maggard M, Agustin M. Quality in surgery: current issues for the future. *World J Surg*. 2005;29:1204–1209.
23. Gerteis M, Edgman-Levitan S, Daley J, et al. Through the patient's eyes: understanding and promoting patient-centered care. *J Healthc Qual*. 1997;19:43.
24. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19:539–549.
25. Merkow RP, Massarweh NN. Looking beyond perioperative morbidity and mortality as measures of surgical quality. *Ann Surg*. 2022;275:e281–e283.
26. Rechel B, McKee M, Haas M, et al. Public reporting on quality, waiting times and patient experience in 11 high-income countries. *Health Policy*. 2016;120:377–383.
27. Flynn KE, Dombeck CB, DeWitt EM, et al. Using item banks to construct measures of patient reported outcomes in clinical trials: investigator perceptions. *Clin Trials*. 2008;5:575–586.
28. Valderas J, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res*. 2008;17:179–193.
29. Churruca K, Pomare C, Ellis LA, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. *Health Expect*. 2021;24:1015–1024.
30. Sokas C, Hu F, Edelen M, et al. A review of PROM implementation in surgical practice. *Ann Surg*. 2022;275:85–90.
31. Warnakulasuriya SR, Patel RC, Singleton GF, et al. Patient-reported outcomes for ambulatory surgery. *Curr Opin Anesthesiol*. 2020;33:768–773.
32. Tong A, Oberbauer R, Bellini MI, et al. Patient-reported outcomes as endpoints in clinical trials of kidney transplantation interventions. *Transpl Int*. 2022;52:10134.
33. Vedadi A, Khairalla R, Che A, et al. Patient-reported outcomes and patient-reported outcome measures in liver transplantation: a scoping review. *Qual Life Res*. 2023;32:2435–2445.
34. Villa BP, Alotaibi S, Brozzi N, et al. Prognostic value of patient-reported outcome measures in adult heart-transplant patients: a systematic review. *J Patient Rep Outcomes*. 2022;6:23.
35. Yang LY, Manhas DS, Howard AF, et al. Patient-reported outcome use in oncology: a systematic review of the impact on patient-clinician communication. *Support Care Cancer*. 2018;26:41–60.
36. Greenhalgh J, Gooding K, Gibbons E, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. *J Patient Rep Outcomes*. 2018;2:42.
37. Antunes B, Harding R, Higginson IJ, et al. Implementing patient-reported outcome measures in palliative care clinical practice: a systematic review of facilitators and barriers. *Palliat Med*. 2014;28:158–175.
38. Mou D, Sisodia RC, Castillo-Angeles M, et al. The surgeon's perceived value of patient-reported outcome measures (PROMs): an exploratory qualitative study of 5 different surgical subspecialties. *Ann Surg*. 2022;275:500–505.

39. Yeo TP, Fogg RW, Shimada A, et al. The imperative of assessing quality of life in patients presenting to a pancreaticobiliary surgery clinic. *Ann Surg*. 2022;277:e136–e143.
40. Foster A, Croot L, Brazier J, et al. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. *J Patient Rep Outcomes*. 2018;2:1–16.
41. Ortega G, Allar BG, Kaur MN, et al. Prioritizing health equity in patient-reported outcome measurement to improve surgical care. *Ann Surg*. 2022;275:488–491.
42. Bhattad PB, Pacifico L. Empowering patients: promoting patient education and health literacy. *Cureus*. 2022;14:e27336.
43. Neumayer L, Giobbie-Hurder A, Jonasson O, et al. Open mesh versus laparoscopic mesh repair of inguinal hernia. *N Engl J Med*. 2004;350:1819–1827.
44. Neumayer L, Jonasson O, Fitzgibbons R, et al. Tension-free inguinal hernia repair: the design of a trial to compare open and laparoscopic surgical techniques. *J Am Coll Surg*. 2003;196:743–752.
45. McCarthy M Jr, Chang CH, Pickard AS, et al. Visual Analog Scales for assessing surgical pain. *J Am Coll Surg*. 2005;201:245–252.
46. McCarthy M Jr, Jonasson O, Chang CH, et al. Assessment of patient functional status after surgery. *J Am Coll Surg*. 2005;201:171–178.
47. Hawn MT, Itani KM, Giobbie-Hurder A, et al. Patient-reported outcomes after inguinal herniorrhaphy. *Surgery*. 2006;140:198–205.
48. Gero D, Muller X, Staiger RD, et al. How to establish benchmarks for surgical outcomes?: a checklist based on an International Expert Delphi Consensus. *Ann Surg*. 2022;275:115–120.
49. Staiger RD, Schwandt H, Puhan MA, et al. Improving surgical outcomes through benchmarking. *Br J Surg*. 2019;106:59–64.
50. Kim TH, Suh YS, Huh YJ, et al. The Comprehensive Complication Index (CCI) is a more sensitive complication index than the conventional Clavien-Dindo classification in radical gastric cancer surgery. *Gastric Cancer*. 2018;21:171–181.
51. Busweiler LAD, Schouwenburg MG, van Berge Henegouwen MI, et al. Textbook outcome as a composite measure in oesophagogastric cancer surgery. *Br J Surg*. 2017;104:742–750.
52. Rossler F, Sapisochin G, Song G, et al. Defining benchmarks for major liver surgery: a multicenter analysis of 5202 living liver donors. *Ann Surg*. 2016;264:492–500.
53. Abbassi F, Gero D, Muller X, et al. Novel benchmark values for redo liver transplantation: does the outcome justify the effort? *Ann Surg*. 2022;276:860–867.
54. Breuer E, Mueller M, Doyle MB, et al. Liver transplantation as a new standard of care in patients with perihilar cholangiocarcinoma? Results from an International Benchmark Study. *Ann Surg*. 2022;276:846–853.
55. Gero D, Raptis DA, Vleeschouwers W, et al. Defining global benchmarks in bariatric surgery: a retrospective multicenter analysis of minimally invasive Roux-en-Y gastric bypass and sleeve gastrectomy. *Ann Surg*. 2019;270:859–867.
56. Gero D, Vannijvel M, Okkema S, et al. Defining global benchmarks in elective secondary bariatric surgery comprising conversional, revisional, and reversal procedures. *Ann Surg*. 2021;274:821–828.
57. Mueller M, Breuer E, Mizuno T, et al. Perihilar cholangiocarcinoma — novel benchmark values for surgical and oncological outcomes from 24 expert centers. *Ann Surg*. 2021;274:780–788.
58. Muller PC, Breuer E, Nickel F, et al. Robotic distal pancreatectomy, a novel standard of care? benchmark values for surgical outcomes from 16 international expert centers. *Ann Surg*. 2023;278:253–259.
59. Muller X, Marcon F, Sapisochin G, et al. Defining benchmarks in liver transplantation: a multicenter outcome analysis determining best achievable results. *Ann Surg*. 2018;267:419–25.
60. Raptis DA, Linecker M, Kambakamba P, et al. Defining benchmark outcomes for ALPPS. *Ann Surg*. 2019;270:835–841.
61. Raptis DA, Sanchez-Velazquez P, Machairas N, et al. Defining benchmark outcomes for pancreatoduodenectomy with portomesenteric venous resection. *Ann Surg*. 2020;272:731–737.
62. Sanchez-Velazquez P, Muller X, Malleo G, et al. Benchmarks in pancreatic surgery: a novel tool for unbiased outcome comparisons. *Ann Surg*. 2019;270:211–218.
63. Schlegel A, van Reeven M, Croome K, et al. A multicentre outcome analysis to define global benchmarks for donation after circulatory death liver transplantation. *J Hepatol*. 2022;76:371–382.
64. Schmidt HM, Gisbertz SS, Moons J, et al. Defining benchmarks for transthoracic esophagectomy: a multicenter analysis of total minimally invasive esophagectomy in low risk patients. *Ann Surg*. 2017;266:814–821.
65. Staiger RD, Rössler F, Kim MJ, et al. Benchmarks in colorectal surgery: multinational study to define quality thresholds in high and low anterior resection. *Br J Surg*. 2022;109:1274–1281.

Copyright © 2023 Wolters Kluwer Health, Inc. All rights reserved.