Users' guide to evidence-based surgery: how to use an article evaluating surgical interventions

John D. Urschel, MD; Charles H. Goldsmith, PhD; Ved R. Tandan, MD; John D. Miller, MD; for the Evidence-Based Surgery Working Group*

Surgeons have traditionally made therapeutic decisions based on existing surgical dogma, personal experience, recommendations of surgical authorities and thoughtful application of surgical basic sciences. Although this approach has served surgeons and their patients fairly well, ineffective or even harmful treatments can be erroneously accepted as the surgical treatment of choice. Evidence-based surgery emphasizes the need to evaluate properly the efficacy of diagnostic and therapeutic interventions before accepting them as standard surgical practice. Evidence in clinical surgery varies in its quality, and this is reflected in several commonly used grading systems for medical evidence.[1-4] Single case reports represent the lowest level of published evidence. They can be valuable as a stimulus for more formal research or as an important observation about a very rare condition. Retrospective case series are a common form of surgical publication.[5] Comparisons of 2 or more therapies are often attempted in these studies, but the retrospective nature of data collection and difficulties comparing heterogeneous patient populations (often in different treatment eras) limit their usefulness. Prospective nonrandomized studies, such as comparisons of a concurrent cohort of patients, overcome some of the limitations of retrospective data collection and comparison to historical "controls." However, they are prone to bias. Bias is defined as any factor or process that tends to deviate the results or conclusions of a trial systematically (not randomly) away from the truth.[6] A properly designed and conducted randomized controlled trial provides a very high level of evidence on which to base surgical decisions.[1-3] Finally, when several randomized controlled trials exist, a meta-analysis of these trials gives the highest level of evidence to support a specific form of treatment.[3,4] Ideally, surgeons should critically examine published evidence and then adjust their practices accordingly.[6] The purpose of this article is to assist surgeons with this process.

Clinical scenario

You are a surgeon performing an esophagectomy for squamous cell carcinoma of the mid-esophagus. You have selected a laparotomy and right thoracotomy for your operative approach. The operation is going well and you are ready to fashion an esophagogastric anastomosis at the apex of the right hemithorax. You are about to do a hand-sewn anastomosis when your assistant (chief resident) suggests using a stapler. He makes reference to a paper that supposedly shows a reduction in anastomotic leaks with stapled esophagogastric anastomoses. You are convinced by his argument and proceed with a stapled anastomosis using an end-to-end anastomotic stapler.

The patient does well postoperatively apart from some pulmonary atelectasis. Contrast esophagography on postoperative day 7 is satisfactory so oral feeding is slowly commenced. The patient is discharged home on postoperative day 14. Unfortunately, early postoperative dysphagia develops and 3 endoscopic dilatations are needed over the ensuing 3 months. You are frustrated by the anastomotic stricture and wonder if a hand-sewn anastomosis would have been better. You decide to review the literature and base your future surgical interventions on the best possible evidence.

From the Surgical Outcomes Research Centre, McMaster University, Hamilton, Ont.

**See acknowledgements for a listing of the members of the Evidence-Based Surgery Working Group.*

Accepted for publication Aug. 17, 2000.

Correspondence to: *Dr. John D. Urschel, Department of Surgery, St. Joseph's Hospital, 50 Charlton Ave. E, Hamilton ON L8N 4A6; fax 905 521-6190, urschelj@mcmaster.ca*

The search

Your search is designed to include a large number of relevant citations initially, but ultimately to permit a very focused review of key articles.[7] You search MEDLINE and seek to create one large set of articles dealing with esophagectomy surgery and another large set dealing with anastomoses. You search the Medical Subject Headings (MeSH) of "esophagectomy," and "esophagus/surgery" ("esophagus/su"). These citations are combined together using the Boolean "OR" function. Similarly, searching the textword "anastomosis" ("anastomosis.tw") creates a large set of articles related to anastomoses. Using the text word instead of the MeSH keyword for anastomosis ensures inclusion of abstracts containing this text word, even if the indexers neglected to index the article for the subject of anastomoses. The 2 large sets ("esophagectomy" OR "esophagus/su," "anastomosis.tw") are then combined using the Boolean "AND" function. For practical reasons the combined set is now limited to the English-language, although this common strategy does eliminate good studies and can introduce bias into systematic reviews.[8] The resulting set is large and inclusive. It is useful for browsing but it is too large for your current purpose. Finally, the set is limited by publication type. You select "randomized controlled trial" in the publication type menu and find less than 10 articles. Four of these compare hand-sewn and stapled anastomoses.[9–12] Although your search successfully yields 4 randomized controlled trials for review, it is important to realize that some trials may have escaped detection during the search process. The indexing of trials is less than perfect so use of additional publication types ("clinical trial, phase III" or "controlled clinical trial") in the publication types menu may increase sensitivity.[5,7] In the library you retrieve the 4 randomized controlled trials and some other articles that caught your eye during MEDLINE browsing.

A review of the various prospective cohort and retrospective comparative studies suggests that stapled esophagogastric anastomoses have a lower rate of leakage than hand-sewn anastomoses, but they may be associated with a higher incidence of anastomotic strictures.[13–20] However, you note that many of these studies compared contemporary stapled anastomotic experience with earlier hand-sewn experience. These papers exemplify the problem of using historical controls in surgical research.[21] You review the 4 randomized controlled trials carefully.[9–12] One is a multicentre study that permitted a wide variety of approaches to esophagectomy,[9] another is a single institution study using a left thoracoabdominal esophagectomy,[10] and the third is a small study using a left cervicotomy for anastomosis.[11] The last paper is a single institution study using the Lewis (laparotomy and right thoracotomy) esophagectomy.[12] Since the Lewis esophagectomy is typically the operative approach that you use, this last paper, by Law and associates, is most applicable to your practice. Your goal is to determine whether this paper should influence the care of your patients. You critique it with this goal in mind.

Critique of an article

Three questions should be answered when critiquing an article comparing surgical interventions: Are the results valid? What are the results? Are the results applicable to my patients?[22–24] (Table 1). These 3 questions will be discussed using the article of Law and associates (Table 2) as an example.

Table 1

How to Critique an Article Evaluating Surgical Interventions

Question	Critique
Are the results valid?	• Was patient assignment randomized, and the randomization process "concealed"? • Were all patients who entered the trial accounted for? • Was follow-up adequate? Were patients analyzed according to the "intention to treat" principle? • Were study personnel "blinded" to treatment? • Were the patient groups similar before treatment? • Apart from the experimental intervention, were the groups treated equally?
What are the results?	• How large was the treatment effect? • How precise was the estimate of the treatment effect?
Are the results applicable to my patients?	• Were the study patients similar to my patients? • Were the measured outcomes clinically relevant? • Are my surgical skills similar to those of the study surgeons?

Table 2

Data Used in This Review, Adapted From the Article by Law and Associates[12]

Data	Type of anastomosis		*p* value
	Hand-sewn	Stapled	
Patients , no.	61	61	
Anastomotic leaks, no (and %)	1 (1.6)	3 (4.9)	0.31
Deaths within 30 d, no. (and %)	0 (0.0)	3 (4.9)	0.08
Anastomotic strictures, no./no. pts. (and %)*	5/55 (9.1)	20/50 (40.0)	0.0003
Anastomotic strictures, no./no pts. (and %)†	5/61 (8.2)	20/61 (32.8)	0.0008

*Excluding early postoperative deaths, malignant strictures and patients with leaks.
†Includes all randomized patients.

Are the results valid?

If a study comparing surgical interventions is not methodologically sound, we should be cautious about using its recommendations in our practices. The most basic (but not the only) issue of study methodology is the randomization of study participants.[6] The surgical literature is replete with reports of successful operations that have subsequently proven to be ineffective or even harmful when examined with randomized controlled trials. The extracranial–intracranial bypass trial is just one example of this phenomenon.[25] The bypass operation made sense from anatomic and physiologic perspectives, and nonrandomized observational studies suggested some benefit for patients with cerebral ischemia. However, the randomized controlled trial showed that the operation was actually harmful to patients. This article highlights the importance of an evidence-based approach to surgical treatments.

Surgical outcomes are influenced not only by the specific surgical intervention but also by patient selection, comorbidities, known prognostic factors and a host of unknown factors. Comparisons of transhiatal and transthoracic esophagectomies, for example, are commonly hampered by an imbalance of chronic lung disease (a comorbidity) in the treatment arms. Randomization is the best way to evenly distribute these determinants of outcome between the treatment and control groups. In addition, the randomization process must be "concealed" from investigators (i.e., the investigator cannot anticipate the direction of randomization for any given patient) until after the patient is entered into the study.[26] If an investigator knows a patient's random allocation destiny before entering the patient on the trial, bias will likely be introduced and the benefits of randomization will be negated. Law's study used intraoperative randomization by a closed envelope method. The closed envelope suggests that the randomization process was concealed, but the authors could have stated this more explicitly in the article.

Readers should be concerned if all patients entered in a trial are not accounted for at the conclusion. Trials may have some patients drop out, and a few patients may be excluded after randomization because of protocol violations. However, these patients must be detailed in the report so the reader can be assured that this was a random, as opposed to a systematic, occurrence. Otherwise we should be concerned that patients had dropped out because of treatment morbidity. Similarly, if patients are not evaluated at a consistent time after treatment or duration of follow-up is inadequate we have an incomplete view of all possible outcomes. Patients may manifest complications at various times after treatment so follow-up must be rigorous and consistent. In Law's article every patient was accounted for and follow-up was complete. The mean follow-up was approximately 20 months. This is adequate for the outcomes of interest (leaks, strictures, operative death).

The "intention-to-treat" concept is important in randomized trials. It is not unusual for patients to be randomized to a given treatment, and then either have no treatment at all or have treatment in the other arm of the study. For example, if a randomized trial compared surgery and radiotherapy for a given cancer, some of the patients randomized to surgery would have their surgery cancelled because of deteriorating fitness or a change of attitude regarding the invasive treatment. Similarly, some patients randomized to radiotherapy may decide that surgery is "better" and undergo an operation outside the study protocol. At first glance it seems appropriate to assess patients based on the treatment that they actually received. However, this can introduce bias. For example, if the surgical arm of a trial is purged of the patients who ultimately did not undergo surgery, the surgical treatment may appear to be superior simply because patients destined to do poorly (surgery cancelled) were excluded. If a trial has too many patients who fail to receive their randomly allocated treatment, the conclusions can become nonsensical. For the most part, however, the intention-to-treat strategy is sound; it is based on the notion that confounding variables will balance out and the randomization will provide a valid comparison of the groups. In the study of Law and associates, the "treatment" (anastomotic technique) was delivered within minutes of randomization so we can assume that patients underwent the treatment they were randomized to receive. However, the article makes no mention of incidents of intraoperative stapler malfunction and resulting hand-sewn anastomotic correction. If this happened, the patient should have remained within the stapled group based on the intention-to-treat principle. Since the article does not discuss this issue we are left wondering about an occasional intention-to-treat violation.

In drug trials patients, clinicians, outcome assessors and investigators are usually "blind" to the treatment that individual patients receive. This is done to prevent expectations of treatment outcome from influencing actual outcomes or outcome evaluations. Although blinding is simple in drug trials, it is usually not possible to blind everyone involved in surgical trials. For example, in Law's study patients with stapled anastomoses would be identifiable by visible staples on radiographs or at endoscopy. To minimize bias when blinding is not possible, post-surgical treatment evaluation can be done by a separate group of clinicians; in theory these clinicians have no interest in the trial outcome. In Law's study, for example, a separate team of clinicians could have done follow-up clinic evaluations (assessment of symptomatic dysphagia)

and endoscopies (assessment of possible strictures) to reduce the possibility of bias. This does not appear to be the case and therefore bias in the nonblinded evaluation is possible.

If a study is sufficiently large, and randomized, the distribution of patients with various prognostic factors should be "balanced" in the treatment and control groups. However, randomized trials with small patient numbers (less than 30 per group) run the risk of comparing, by chance, 2 fairly dissimilar groups of patients. Readers can be reassured that chance imbalance of the 2 groups does not occur if the known prognostic variables are detailed and the issue of chance imbalance is discussed. In Law's article data are given on age, sex, anatomic location of tumour, stage of tumour and size of esophagus. The patient groups appear very similar before treatment. Similarly, the reader should be wary of differences in general treatment of the 2 study groups. These differences are called "cointerventions." If, for example, minimally invasive surgery were to be compared to open surgery, provision of epidural analgesia in only 1 arm of the study would be an important cointervention that could bias the results. In Law's study the patients were treated identically except for the experimental intervention itself.

What are the results?

Surgical randomized trials commonly measure dichotomous outcomes or events such as death, cancer recurrence and surgical complications. These dichotomous events either happen or they don't, so the article usually reports the proportion of patients having the event of interest. Analysis and presentation of results is therefore a process of comparing proportions. These comparisons can be made and expressed in various ways; readers should be familiar with the various methods and terms (Table 3). In the article of Law and associates, for example, strictures developed in 9% of the patients with hand-sewn anastomoses and 40% of patients with stapled anastomoses (Table 2). The absolute risk reduction for the hand-sewn anastomotic technique is 0.31 (0.40 – 0.09 = 0.31). In other words, the hand-sewn technique resulted in a 31% absolute reduction in anastomotic strictures. This is one way of expressing the results. Another way is to give the relative risk: the risk of events in one group relative to that in the other group. In Law's article the relative risk of anastomotic stricturing for the hand-sewn technique is 0.22 (0.09/0.40 = 0.22). In other words, the occurrence of strictures with the hand-sewn technique was about one-fifth (22%) of that seen with the stapled technique. The complement of relative risk, termed relative risk reduction, is a common way to express results of dichotomous outcomes. Since it is the complement of relative risk, it is calculated by subtracting relative risk from 1. By convention it is expressed as a percentage. The relative risk reduction for anastomotic stricturing with the hand-sewn technique is 78% (1 – 0.22 = 0.78 × 100 = 78%). One could say that using the hand-sewn technique reduced the risk of strictures by 78%.

Randomized surgical trials study a treatment in a small group of patients and then provide an estimate of treatment effect that is applicable to the larger population of patients we treat. That raises an important question: How precise is the estimate of treatment effect? Investigators can indicate the precision of their treatment estimate in several ways. One approach is to give the traditional p value. Law and associates used this approach when giving the anastomotic stricture results. The p value for the comparison of stricture incidence in the 2 treatment groups was 0.0003 (χ^2 test). This means that the chance of the observed difference in outcome being a random event (when there is no true difference in the 2 procedures) is 3 in 10 000. Of note, the conventional criterion for p values for "positive" studies in medicine is 0.05 (chance of the observed difference being a random event is 5 in 100). This particular threshold for statistical significance is completely arbitrary, so investigators should report specific p values. Statements such as "$p < 0.05$, significant" or "$p > 0.05$, not significant" should not be used. Finally, readers should be aware that 2 treatments may be significantly different statistically without this difference being clinically important.

Although Law's provision of the p value reassures us that the reported difference is in fact statistically signifi-

Table 3

Terms Used to Show the Magnitude and Precision of the Treatment Effect (Decreased Risk of Anastomotic Stricture With the Hand-Sewn Technique)*

Term	Description	Example
Absolute risk reduction	Absolute reduction in events in one group compared with the other	40% – 9% = 31% absolute risk reduction for stricture with the hand-sewn technique
Relative risk	Risk of events in one group relative to the other	9% ÷ 40% = 0.22 relative risk of stricture with the hand-sewn technique
Relative risk reduction	Complement of relative risk, expressed as a percentage	1 – 0.22 = 0.78 × 100 = 78% reduction in stricture risk with the hand-sewn technique
95% confidence interval	An interval of values that include the true value 95% of the time (calculated)	78% (CI 47% to 90%) reduction in stricture risk with the hand-sewn technique

*Data used in the table were taken from the article by Laws and associates.[12]

cant, there are other ways of communicating this information. The confidence interval is another, and often preferable way, of showing the precision of the treatment effect estimate. Traditionally, a 95% confidence interval is reported. These confidence intervals define an interval of values that include the true value 95% of the time. Law and associates did not provide confidence intervals, but we are able to calculate them ourselves since the raw data were provided.[27] The relative risk reduction for anastomotic stricturing with the hand-sewn technique is 78%, with a 95% confidence interval from 47% to 90%. We can see that a relative risk reduction anywhere within this interval is clinically important. We are therefore confident that the estimate of treatment effect (relative risk reduction for stricturing of 78% with the hand-sewn technique) is sufficiently precise to accept it. Having determined the magnitude and precision of a treatment effect, the next question involves the issue of general applicability of the results to a larger patient population.

Are the results applicable to my patients?

Surgeons should ask this question before incorporating trial recommendations into their clinical practices. If a randomized trial had very rigid inclusion criteria for enrolment, the results may not be applicable to the larger population of surgical patients[28,29] The Veterans Affairs Gastroesophageal Reflux Study Group, for example, showed that fundoplication was superior to medical management for complicated reflux disease in male veterans.[30] Caution should be exercised in generalizing these results to other populations (women, non-veterans and patients with uncomplicated reflux) or to patients being treated with more effective medications. Patient age and fitness are other trial inclusion criteria that pose problems when generalizing results. For example, if a randomized trial only involved fit patients its recommendations for an aggressive treatment strategy would not necessarily be applicable to old or frail patients. Surgeons should examine the inclusion criteria to see if the study results are applicable to their patients. In the article of Law and associates, the enrolment criterion was defined as any patient undergoing Lewis esophagectomy for squamous carcinoma of the esophagus. The results should therefore be applicable to the patient in our clinical scenario.

Before accepting one treatment as superior to another, surgeons should question the clinical relevance of the main outcome measures. For example, randomized trials comparing laparoscopic and open surgery often use length of hospital stay as a major outcome measure. Although length of stay is important, it may not be the most clinically relevant outcome of interest to surgeons or patients. In Law's article the time needed to construct a hand-sewn and a stapled anastomosis was compared. This outcome is not as important as the other outcomes measured, such as operative mortality, anastomotic leakage and anastomotic stricturing. Most surgeons and patients would agree that these are 3 clinically important anastomotic outcome issues.[20,31]

Finally, surgeons should consider their own skill and experience before applying research findings to their practices. Surgeons must ask themselves if they are as technically proficient with the reported operative procedures as the investigating team of surgeons. Alternatively, the investigating surgeons' proficiency with a given operation may be open to question.[32,33] Proficiency with a particular operation (on the part of an individual surgeon or a team of surgical investigators) is obviously a very important component of its effectiveness.[34] If published evidence conclusively favours an operation that a surgeon does not perform, or does not perform well, that surgeon is faced with 3 choices: proceed with another operation, refer the patient to a colleague or seek additional training to master the operation.[35,36] This issue of surgical proficiency creates a difficult dilemma for practising surgeons and highlights a fundamental difference between medical and surgical trials.[6,37] A lack of surgical proficiency will bias a randomized controlled trial in favour of the simpler of 2 operative procedures. On the other hand, a randomized controlled trial involving a very complex operation performed by exceptionally skilled surgeons may not be applicable to the larger surgical community.

Resolution of the clinical scenario

A careful critique of the article by Law and associates leads one to conclude that both stapled and hand-sewn esophagogastric anastomoses are acceptable in terms of the most critical immediate outcomes (death and anastomotic leaks). However, anastomotic stricturing causes dysphagia and impairs quality of life.[38,39] Since palliation of dysphagia is one of the goals of esophagectomy for cancer, postoperative anastomotic stricturing is a major concern. Hand-sewn anastomoses carry a lower risk of anastomotic stricture. As the surgeon in our clinical scenario, you decide to revert to your previous practice of constructing esophagogastric anastomoses with a hand-sewn technique.

Acknowledgements: The Evidence-Based Surgery Working Group members are as follows: Stuart Archibald, MD;*†‡ Mohit Bhandari, MD;† Charles H. Goldsmith, PhD;‡§ Dennis Hong, MD;† John D. Miller, MD;*†‡ Marko Simunovic, MD, MPH;†‡ Ved Tandan, MD, MSc;*†‡§ Achilleas Thoma, MD;*†‡ John D. Urschel, MD;*†‡ Susan Dimitry, BA.‡ *Department of Surgery, St. Joseph's Hospital, †Department of Surgery, McMaster University, ‡Surgical Outcomes Research Centre, McMaster University, and §Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.