Appropriateness Criteria to Assess Variations in Surgical Procedure Use in the United States

Elise H. Lawson, MD, MSHS; Melinda Maggard Gibbons, MD, MSHS; Angela M. Ingraham, MD; Paul G. Shekelle, MD, PhD; Clifford Y. Ko, MD, MS, MSHS

Objectives: To systematically describe appropriateness criteria (AC) developed in the United States for surgical procedures and to summarize how these criteria have been applied to identify overuse and underuse of procedures in US populations.

Data Sources: MEDLINE literature search performed in February 2010 and May 2011.

Study Selection: Studies were included if they addressed the appropriateness of a surgical procedure using the RAND-UCLA Appropriateness Method. Non-US studies were excluded.

Data Extraction: Information was abstracted on study design, surgical procedure, and reported rates of appropriate use, overuse, and underuse. Identified AC were cross-referenced with lists of common procedures from the Nationwide Inpatient Sample and the State Ambulatory Surgery databases.

Data Synthesis: A total of 1601 titles were identified; 39 met the inclusion criteria. Of these, 17 developed AC and 27 applied AC to US populations. Appropriateness criteria have been developed for 16 surgical procedures. Underuse has only been studied for coronary artery bypass graft surgery, and rates range from 24% to 57%. Overuse has been more broadly studied, with rates ranging from 9% to 53% for carotid endarterectomy, 0% to 14% for coronary artery bypass graft, 11% to 24% for upper gastrointestinal tract endoscopy, and 16% to 70% for hysterectomy. Appropriateness criteria exist for 10 of the 25 most common inpatient procedures and 6 of the 15 top ambulatory procedures in the United States. Most studies are more than 5 years old.

Conclusions: Most existing AC are outdated, and AC have never been developed for most common surgical procedures. A broad and coordinated effort to develop and maintain AC would be required to implement this tool to address variation in the use of surgical procedures.

Arch Surg. 2011;146(12):1433-1440

The US health care system is increasingly focused on promoting patient-centered care, improving clinical outcomes, and reducing variations in the provision of care. For surgery, the goal of these quality improvement efforts is to perform the right procedure for the right patient in the right way.

Considerable evidence shows widespread variations in rates of surgical procedures in the United States. Well-documented racial disparities and geographic variations in the use of surgical procedures are not fully explained by disease incidence or patient preferences. These findings are more likely the result of overuse and underuse of surgical procedures.

Recent commentaries have proposed using the RAND-UCLA Appropriateness Method (RUAM) to improve the quality of surgical care by addressing variations in the use of surgical procedures. The RUAM synthesizes the best available evidence with expert clinical judgment to produce appropriateness criteria (AC), which weigh the relative risks and benefits of a procedure for specific clinical scenarios. These AC can be applied to patient populations to assess for overuse and underuse.

The purposes of this study are to systematically describe AC developed in the United States for surgical procedures and to summarize how these criteria have been applied to identify overuse and underuse. Our goals were to elucidate what is known and not known about the appropriate use of common inpatient and ambulatory procedures in the United States and to determine whether there is a need for future development and application of AC.
An indication is considered necessary if it would be implemented if the procedure is worth doing, exclusive of cost.7(p55) The magnitude of the benefit is not small.8

Indications are classified in 1 of 3 categories: appropriate, equivocal, and inappropriate. An appropriate indication for a procedure is one for which “the expected health benefit (eg, increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeds the expected negative consequences (eg, mortality, morbidity, anxiety, pain, time lost from work) by a sufficiently wide margin that the procedure is worth doing, exclusive of cost.”7(p55) This method starts with an extensive review of the literature on the risks and benefits of the procedure. A comprehensive and mutually exclusive set of clinical scenarios or indications for the procedure is then compiled, complete with specific definitions for any potentially ambiguous terms (eg, failed medical therapy would be explicitly defined). Because of the need to be all inclusive, the list typically includes many hundreds of specific clinical circumstances.

An expert panel then rates each indication in 2 rounds, with the second round occurring after an in-person discussion of the first-round results. Indications are classified as “appropriate” (the expected benefits of the procedure outweigh the expected harms), “equivocal” (the expected benefits and harms are roughly equal or there is disagreement amongst the panelists), or “inappropriate” (the expected harms outweigh the expected benefits). Appropriate indications are sometimes further classified as “necessary” by the panel, usually in a third round. An indication is considered necessary if it would be improper care to not offer the procedure to the patient, there is a reasonable chance the procedure will benefit the patient, and the magnitude of the benefit is not small.8 Table 1 lists examples of indications with each of these classifications.

Overuse and underuse of a surgical procedure are determined by applying these theoretical indications to actual patients. Underuse is defined as any patient with a necessary indication who does not receive the procedure. The study sample is derived from a pool of patients who may or may not undergo the procedure. For example, patients who are diagnosed with coronary artery disease after undergoing coronary angiography then proceed with either coronary revascularization or medical management. Overuse is defined as any patient who undergoes a procedure for an inappropriate indication.

The study sample for overuse (and for appropriate use) is thus derived from patients who underwent the procedure.

A substantial amount of research has been performed regarding the reliability and validity of the RUAM. Studies of test-retest reliability of the AC for coronary revascularization have been demonstrated,11 and there is evidence supporting the predictive validity of the AC for coronary revascularization.17-19 For example, a prospective study19 on the appropriate use of coronary angiography then proceed with either coronary revascularization or medical management. Underuse is defined as any patient with a necessary indication who does not receive the procedure. The study sample is derived from a pool of patients who may or may not undergo the procedure. For example, patients who are diagnosed with coronary artery disease after undergoing coronary angiography then proceed with either coronary revascularization or medical management. Overuse is defined as any patient who undergoes a procedure for an inappropriate indication.

The RUAM was developed to systematically assess variation in the use of surgical procedures by defining which patients should and should not undergo surgical intervention vs medical therapy. An appropriate indication for a procedure is one for which “the expected health benefit (eg, increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeds the expected negative consequences (eg, mortality, morbidity, anxiety, pain, time lost from work) by a sufficiently wide margin that the procedure is worth doing, exclusive of cost.”7(p55) This method starts with an extensive review of the literature on the risks and benefits of the procedure. A comprehensive and mutually exclusive set of clinical scenarios or indications for the procedure is then compiled, complete with specific definitions for any potentially ambiguous terms (eg, failed medical therapy would be explicitly defined). Because of the need to be all inclusive, the list typically includes many hundreds of specific clinical circumstances.

An expert panel then rates each indication in 2 rounds, with the second round occurring after an in-person discussion of the first-round results. Indications are classified as “appropriate” (the expected benefits of the procedure outweigh the expected harms), “equivocal” (the expected benefits and harms are roughly equal or there is disagreement amongst the panelists), or “inappropriate” (the expected harms outweigh the expected benefits). Appropriate indications are sometimes further classified as “necessary” by the panel, usually in a third round. An indication is considered necessary if it would be improper care to not offer the procedure to the patient, there is a reasonable chance the procedure will benefit the patient, and the magnitude of the benefit is not small.8 Table 1 lists examples of indications with each of these classifications.

Overuse and underuse of a surgical procedure are determined by applying these theoretical indications to actual patients. Underuse is defined as any patient with a necessary indication who does not receive the procedure. The study sample is derived from a pool of patients who may or may not undergo the procedure. For example, patients who are diagnosed with coronary artery disease after undergoing coronary angiography then proceed with either coronary revascularization or medical management. Overuse is defined as any patient who undergoes a procedure for an inappropriate indication.
DATA ABSTRACTION

Study results were abstracted into data tables. Data abstracted included dates of study and publication, surgical procedure, details of the study design, and reported rates of appropriate use, overuse, and underuse.

Studies that developed AC are reported if they included comprehensive and mutually exclusive varying clinical scenarios rated by a panel in at least 2 rounds, with at least 1 round occurring after an in-person discussion. We excluded guidelines and consensus statements that were not developed using the RUAM. We also excluded studies that used a non-US panel. Studies assessing for appropriate use, overuse, and/or underuse of a surgical procedure were included if they applied AC developed by a US panel to a random sample or stratified sample of the US population.

IDENTIFICATION OF COMMON INPATIENT AND AMBULATORY SURGICAL PROCEDURES

To identify what is known and not known about the appropriateness of surgical care in the United States, we identified the most commonly performed inpatient and ambulatory procedures and cross-referenced these lists with the results of our literature search. The Nationwide Inpatient Sample, a Healthcare Cost and Utilization Project data set, was used to identify the most frequently performed inpatient surgical procedures in 2008 and the associated mean costs for these procedures. The 2008 Nationwide Inpatient Sample is the latest year for which summarized data were available and is a 20% stratified sample of US community hospitals. This data set contains all discharge data on 8 million hospitalizations from 1056 hospitals located in 42 states. International Classification of Diseases, Ninth Revision (ICD-9) codes for procedures were grouped into clinically meaningful categories using the Clinical Classification Software (http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp).

The State Ambulatory Surgery Databases, a set of Healthcare Cost and Utilization Project data sets from 28 participating states, were used to identify the most frequently performed ambulatory surgical procedures in 2007 and the associated mean charges for these procedures. The 2007 State Ambulatory Surgery Databases are the latest year for which summarized data are available. These databases capture discharge data for surgical procedures performed on the same day in which patients are admitted and released. Some of the databases contain the ambulatory surgery encounter abstracts for that state (including records from both hospital-affiliated and freestanding surgery centers). Procedures were grouped by the Clinical Classification Software, as was done with the inpatient procedures.

Lists of common inpatient and ambulatory surgical procedures were created using these sources. Procedures that are not commonly performed by surgeons were excluded (eg, bronchoscopy and percutaneous coronary angioplasty), as were procedures related to pregnancy or childbirth (eg, cesarean section and circumcision), relatively low-risk procedures (eg, incision and drainage of skin lesions and wound debridement), and bedside procedures (eg, thoracentesis and abdominal paracentesis). Some categories defined by the Clinical Classification Software were combined to form new categories (low back surgery combines spinal fusion and laminectomy, excision intervertebral disk; transurethral prostatectomy is combined with excision, drainage, or removal of urinary obstruction). Other overly broad categories were excluded (eg, other operating room upper gastrointestinal [GI] therapeutic procedures). Because bariatric procedures are part of this broad category, a different source was used to determine the frequency and cost of these procedures.24

RESULTS

DESCRIPTION OF STUDIES IDENTIFIED BY THE LITERATURE SEARCH

Our search identified 1601 articles, of which 395 were screened and of which 39 were included in this review (Figure 1). Articles were excluded if they did not address the appropriateness of a surgical procedure using AC rated by a US panel using the RUAM.

Of the included articles, 12 developed AC and 22 applied AC to assess for appropriate use, overuse, and/or underuse. In addition, 5 studies both developed and applied AC. Some articles addressed more than 1 surgical procedure. Articles that evaluate the reliability and validity of the RUAM will be summarized elsewhere.

STUDIES ON DEVELOPMENT OF AC

In the United States, AC have been developed for 16 surgical procedures: coronary artery bypass grafting (CABG), 10,12,25,26 bariatric procedures, 9 abdominal aortic aneurysm surgery, 7,28 carotid endarterectomy, 20,28,29 hysterectomy, 12,36 nephrectomy for metastatic renal cell cancer, 11 cholecystectomy, 26,34 upper GI tract endoscopy, 20,32 colonoscopy, 20,33 cataract procedures, 28 tympanostomy tube placement, 33 sinus procedures, 36 low back sur-
gery for sciatica, sentinel lymph node biopsy for melanoma, and tonsillectomy or adenoidectomy, and carpal tunnel surgery. For some procedures, AC have been developed multiple times, with CABG being the most frequently studied procedure (4 studies) followed by carotid endarterectomy (3 studies). Most studies were published more than 5 years ago, which is a proposed threshold for updating practice guidelines as new evidence accumulates.

APPLICATION OF AC TO ASSESS FOR APPROPRIATE USE, UNDERUSE, AND OVERUSE

We identified 27 unique studies that applied AC developed by a US panel to a sample of the US population. Coronary artery bypass grafting and carotid endarterectomy were again the most frequently studied (10 and 8 studies, respectively). Most studies were published more than 5 years ago, which is a proposed threshold for updating practice guidelines as new evidence accumulates.

Appropriate use has been assessed for 8 procedures: CABG, carotid endarterectomy, hysterectomy, upper GI tract endoscopy, cataract procedures, tympanostomy tube placement, sinus procedures, and low back surgery for sciatica (Table 2 and Table 3). In contrast, underuse has only been reported in US populations for CABG. Studies analyzing underuse of CABG in varied populations of patients who underwent coronary angiography in the early 1990s report rates ranging from 24% to 25%. In 3 studies, this rate was as high as 42%, and in other studies, as high as 57%. Overuse has been studied for 9 procedures, with rates ranging from 9% to 53% for carotid endarterectomy, 0% to 14% for CABG, 11% to 24% for upper GI tract endoscopy, and 16% to 70% for hysterectomy (Figure 2 and Tables 2 and 3). Procedures for which overuse has only been reported once include cataract procedures (2%), tympanostomy tube placement (23%), sinus procedures (16%), and low back surgery for sciatica (31%).
Table 3. Availability of Appropriateness Criteria and Results of Application for Common Ambulatory Surgical Procedures

<table>
<thead>
<tr>
<th>Rank</th>
<th>Procedure</th>
<th>Ambulatory Procedures, in Thousands</th>
<th>Mean Charge per Visit, $</th>
<th>Appropriateness Criteria Year(s) Developed (Published)</th>
<th>Reports of Application, %</th>
<th>Appropriate Use</th>
<th>Underuse</th>
<th>Overuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Colonoscopy and biopsy</td>
<td>2247</td>
<td>2369</td>
<td>1994, 1984, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Lens and cataract procedures</td>
<td>689</td>
<td>4780</td>
<td>(2008), 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Excision of semilunar cartilage of knee</td>
<td>317</td>
<td>7357</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Cholecystectomy and common duct exploration</td>
<td>313</td>
<td>10,838</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Tonsillectomy and/or adenoidectomy</td>
<td>306</td>
<td>5286</td>
<td>(1994)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Inguinal and femoral hernia repair</td>
<td>267</td>
<td>8187</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Other excision of cervix and uterus</td>
<td>266</td>
<td>7362</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Breast biopsy and other diagnostic procedures on breast</td>
<td>241</td>
<td>3863</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Lumpectomy, quadrantectomy of breast</td>
<td>235</td>
<td>7458</td>
<td>(1994)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Endoscopy and endoscopic biopsy of the urinary tract</td>
<td>199</td>
<td>4300</td>
<td>(1994)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Partial excision, bone</td>
<td>191</td>
<td>7590</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Diagnostic dilatation and curettage</td>
<td>190</td>
<td>6366</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Transurethral excision, drainage, or removal of urinary obstruction</td>
<td>186</td>
<td>8044</td>
<td>1994, 2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aNumber of ambulatory procedures and mean charges per visit from the State Ambulatory Surgery Databases, 2007.

AVAILABILITY AND APPLICATION OF AC FOR COMMON SURGICAL PROCEDURES

Tables 2 and 3 demonstrate the many gaps in what is known regarding the appropriateness of major inpatient and ambulatory procedures, respectively. Appropriateness criteria exist for 10 of the 25 most commonly performed inpatient procedures and have been applied to assess for overuse and/or underuse for 6. Notably, AC have not been developed for half of the 10 most commonly performed inpatient procedures and have only been applied for 3 of these procedures. Some common inpatient procedures for which information regarding appropriateness is lacking include knee arthroplasty, partial and total hip replacement, appendectomy, colorectal resection, and treatment of hip and femur fractures and dislocations.

Less is known regarding the appropriate use of common ambulatory procedures. Of the top 15 ambulatory procedures performed in the United States, AC exist for only 6 of them and have been applied for only 3. Some common ambulatory procedures for which information regarding appropriateness is lacking include excision of semilunar cartilage of the knee, inguinal and femoral hernia repair, excision of cervix and uterus, and breast procedures, including biopsy, lumpectomy, and quadrantectomy.

Appropriateness criteria developed in the United States exist for 16 surgical procedures, and application of these criteria to varied US populations has revealed nontrivial rates of overuse and underuse. However, most AC are more than 5 years old, and furthermore, AC have never been developed for most common inpatient and ambulatory procedures.

Quality improvement efforts for the field of surgery, such as the Surgical Care Improvement Project (SCIP) and the National Surgical Quality Improvement Program (NSQIP), focus exclusively on improving processes and outcomes of care, respectively. Much less effort is spent on improving patient selection before surgery. A national effort to develop AC with clinical buy-in and decision support could improve surgical care by addressing disparities in the receipt of surgical procedures and reducing overuse and underuse. More appropriate patient selection could also potentially improve postoperative outcomes because studies on CABG show that patients who are treated in accordance with their appropriateness category have better clinical outcomes compared with patients who are treated inappropriately or who do not receive necessary interventions.

Developing AC and reporting rates of overuse and underuse may in themselves stimulate improved patient selection for surgery but are unlikely to be sufficient. Some groups have developed computer algorithms based on panel appropriateness ratings, which make AC more feasible for implementation in the clinical realm. With use of health information technology rapidly evolving, such algorithms could be integrated into electronic medical records, thus streamlining assessment of appropriateness of surgery for a patient. Surgeons and referring physicians could use AC as a decision support aid, with the algorithm serving as an electronic second opinion. A study by Junghans et al found that patient-specific appropriateness ratings were more effective than guidelines in changing the way physicians treated angina, which sup-

©2011 American Medical Association. All rights reserved.
ports our hypothesis that clinical use of AC could reduce practice variations and improve the overall appropriateness of surgical care.

Use of AC could improve the current informed consent process by providing an independent assessment of the balance of risks and benefits of a procedure for the patient's specific clinical scenario. Indications classified as equivocal could trigger a multidisciplinary approach to assist patients in deciding whether to undergo surgery or proceed with medical management. Routinely providing this information and service to patients is consistent with the current emphasis on patient-centered care and would not aim to replace a physician's clinical judgment but rather to augment or guide decisions.

Administrators or health agencies seeking to reduce practice variation could use AC as a quality metric. They could be implemented as either a process measure (ie, were AC used in the decision-making process or was the patient provided information on appropriateness before surgical intervention) or as an outcome measure (ie, rates of overuse and underuse) for this purpose.

It is unclear what effect more appropriate care would have on cost to the health care system because the balance of overuse and underuse of procedures is unknown. The Health Services Utilization Study demonstrated that differences in overuse did not explain geographic variations in the use of 3 procedures (coronary angiography, carotid endarterectomy, and upper GI tract endoscopy). The researchers assigned appropriateness ratings to patients in areas of high, average, and low use of these procedures and found that the differences in levels of appropriate and inappropriate use among sites were small.33 Reducing variation in the use of surgical procedures is thus not isolated to reducing overuse but also must focus on identifying and reducing underuse.

Comparative effectiveness research is proposed as a means of reducing practice variation; however, the results of these studies will likely not be implemented for many years. In contrast, broad development and implementation of AC for common surgical procedures could occur during a relatively short period. The use of AC is an intermediate step and could set the agenda for future comparative effectiveness research by identifying gray areas where not enough is known regarding the balance of risks and benefits of a procedure (ie, equivocal indications).

Our study has possible limitations. First, studies on the development and application of AC may have been overlooked. To reduce this possibility, we started with broad search terms and had 2 physicians perform the screening and reference mining. Second, the implications of our findings may be limited by the possibility that the RUAM may have differing reliability and validity for different procedures. Most of the methodologic studies on the RUAM focus on a relatively small number of procedures (eg, CABG, carotid endarterectomy, and hysterectomy). Further evaluation of the method is warranted and could be performed concurrently with further development, application, and implementation of AC for a broad range of procedures.

Our initial search produced additional articles that developed AC using non-US panels, which we did not report. Studies comparing the results of AC developed in different countries have reported small but not insignificant differences in appropriateness ratings. A study comparing the appropriateness of CABG use in the United States and Canada using AC developed independently in each country found that 6% of US cases and 4% of Canadian cases were rated as inappropriate by Canadian criteria compared with 2% and 3%, respectively, using US criteria.61 Studies that focused on low back surgery and colonoscopy had similar findings.73,77 Although the differences are small, we believe that using only AC developed by US panels enhances the reliability and validity of the RUAM for detecting overuse and underuse.

In conclusion, AC are a feasible, realistic, and practical tool for addressing the persistent problem of variations in the use of surgical procedures. Apart from CABG, bariatric surgery, and carpal tunnel surgery, most AC are outdated. A coordinated effort to produce and maintain up-to-date AC for surgical procedures that are frequently performed, have elevated risk of morbidity and

Figure 2. Summary of studies on rates of overuse of the following surgical procedures in US populations: carotid endarterectomy,41-42,47-53 coronary artery bypass grafting,25-45 upper gastrointestinal tract endoscopy,41-43 and hysterectomy.12,30,44 Asterisk indicates that year of publication is listed if the year the procedure was performed was not reported. n refers to number of patients studied.
mortality, are controversial, and/or that use significant resources could potentially greatly improve the quality of surgical care if these criteria are integrated into everyday surgical practice. Some professional societies, such as the American College of Cardiology and the American College of Radiology, have already undertaken such programs. Surgical societies could benefit from such an initiative as well.

Accepted for Publication: March 9, 2011.

Correspondence: Elise H. Lawson, MD, MSHS, Department of Surgery, UCLA Medical Center, CHS 72-215, 10833 LeConte Ave, Los Angeles, CA 90095 (elawson@mednet.ucla.edu).

Author Contributions: Study concept and design: Lawson, Gibbons, Ingraham, Shekelle, and Ko. Acquisition of data: Lawson, Gibbons, and Ingraham. Analysis and interpretation of data: Lawson, Gibbons, Ingraham, Shekelle, and Ko. Drafting of the manuscript: Lawson and Gibbons. Critical revision of the manuscript for important intellectual content: Lawson, Gibbons, Shekelle, and Ko. Administrative, technical, or material support: Lawson and Gibbons. Study supervision: Gibbons, Shekelle, and Ko. Financial Disclosure: None reported.

Funding/Support: Dr Lawson’s time was supported by the Robert Wood Johnson Foundation Clinical Scholars Program through the American College of Surgeons.

REFERENCES

38. Dubois RW, Swetter SM, Atkins M, et al. Developing indications for the use of...


