The appropriateness method has acceptable reliability and validity for assessing overuse and underuse of surgical procedures

Elise H. Lawson, Melinda Maggard Gibbons, Clifford Y. Ko, Paul G. Shekelle

Department of Surgery, David Geffen School of Medicine, University of California, 10833 LeConte Ave., CHS 72-215, Los Angeles, CA 90095, USA
Division of Research and Optimal Patient Care, American College of Surgeons, 633 N. Saint Clair St., Chicago, IL 60611, USA
Department of Surgery, Olive View-UCLA Medical Center, 14445 Olive View Drive, Sylmar, CA 91342, USA
Departments of Surgery and Medicine, VA Greater Los Angeles Healthcare System, 11301 Wilshire Blvd., Los Angeles, CA 90073, USA
RAND Corporation, 1776 Main St., Santa Monica, CA 90401, USA

Accepted 8 July 2012

Abstract

Objective: To summarize the findings of methodological studies on the RAND/University of California Los Angeles (RAND/UCLA) appropriateness method, which was developed to assess if variation in the use of surgical procedures is because of overuse and/or underuse.

Study Design and Setting: A MEDLINE literature search was performed. Studies were included if they assessed the reliability or validity of the RAND/UCLA appropriateness method for a surgical procedure or the effect of altering panelist composition or eliminating in-person discussion between rating rounds. Information was abstracted on procedure, study design, and findings.

Results: One thousand six hundred one titles were identified, and 37 met the inclusion criteria. The test-retest reliability is good to very good (kappa, 0.64–0.81) for total knee and hip joint replacement, coronary artery bypass grafting (CABG), and carotid endarterectomy (CEA). The interpanel reliability is moderate to very good (kappa, 0.52–0.83) for CABG and hysterectomy. Construct validity has been demonstrated by comparing the appropriateness method with guidelines and/or evidence-based approaches for endoscopy, colonoscopy, CABG, hysterectomy, and CEA. Predictive validity has been studied for cardiac revascularization, in which concordance with appropriateness classification is associated with better clinical outcomes.

Conclusion: Our findings support use of the appropriateness method to assess variation in the rates of the procedures studied by identifying overuse and underuse. Further methodological research should be conducted as appropriateness criteria are developed and implemented for a broader range of procedures. © 2012 Elsevier Inc. All rights reserved.

Keywords: Appropriateness; Surgery; Reliability; Validity; Overuse; Underuse

1. Introduction

Prior reports have demonstrated substantial variations in the rates of surgical procedures performed in the United States that are not fully explained by disease incidence or patient preferences [1–4]. The multifactorial causes of these variations are not fully known but are likely in part related to differences in clinician opinion regarding treatment options [5,6]. Although randomized controlled trials are considered the gold standard for determining the safety and effectiveness of a procedure, the generalizability of results to a broad population is often limited by the enrollment of a narrow spectrum of patients, and the time lag between study planning and dissemination of results can be substantial. Furthermore, relying on trials to determine the best treatment for a wide range of clinical scenarios would be impractical. Consider, for example, designing a trial where four clinical variables are judged as important in deciding whether a patient should undergo surgery (e.g., severe disease vs. mild disease, high surgical risk vs. low surgical risk, prior response to medical therapy, and the presence or absence of only a single comorbid condition). These four clinical variables can define 16 potentially distinct patient populations (2^4 = 16 clinical scenarios). If the study requires 200 people in each comparison group to have sufficient power to detect a moderate-sized difference, this would require the enrollment of 200 × 16 = 3,200 people in the trial. Undertaking and completing randomized controlled trials enrolling 3,200 people for each of the 20 most common surgical procedures would require the infrastructure and funding to enroll, collect data on, and follow-up 64,000 patients. This is not likely to be
receive a procedure that is proven effective and would im-
weigh the benefits) and/or underuse (patient does not
(patient undergoes a procedure for which the risks out-
plications to determine if variation is the result of overuse
appropriateness criteria, can then be applied to patient po-
own clinical judgment. The rated indications, or
ratings based on the evidence in the literature and his/her
exhaustive list of specific clinical indications and assign
indications for the procedure is then compiled, complete with
specific and mutually exclusive set of clinical scenarios or
ature on risks and benefits of the procedure. A comprehen-
sive and mutually exclusive set of clinical scenarios or
indications for the procedure is then compiled, complete with
specific definitions for any potentially ambiguous terms
(e.g., “failed medical therapy” would be explicitly defined).
Because of the need to be inclusive, the list typically includes
many hundreds of clinical indications.
Individuals comprising an expert panel rate each indica-
tion in two 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 among the panelists), or “inappropriate”
(the expected harms outweigh the expected benefits). Ap-
propriate indications are sometimes further classified as
“necessary” by the panel, usually in a third round. An in-
dication is considered necessary if it would be improper
care to not offer the procedure, there is a reasonable chance
the procedure will benefit the patient, and the magnitude of
the benefit is not small [7]. Table 1 lists examples of indi-
cations with each of these classifications [10,11].
Overuse and underuse of a surgical procedure are deter-
mained by applying the aforementioned indications to actual

What is new?

Key findings
- The RAND/University of California Los Angeles (RAND/UCLA) appropriateness method has moderate to very good reliability for determining over-
use and underuse of the surgical procedures studied: total knee and hip joint replacement, cor-
orary artery bypass grafting (CABG), carotid end-
arterectomy (CEA), and hysterectomy.
- The construct validity of appropriateness criteria has been demonstrated for upper gastrointestinal endoscopy, colonoscopy, CABG, hysterectomy, and CEA through comparisons with professional society guidelines and/or evidence-based approaches.
- Concordance with appropriateness criteria classification is associated with better clinical outcomes for cardiac revascularization.

What this adds to what was known?
- Systematic review of methodological studies on the RAND/UCLA appropriateness method.

What is the implication and what should change now?
- Our study supports the use of the RAND/UCLA appropriateness method to assess variation in the use of surgical procedures by identifying overuse and underuse for the procedures studied and highlights the need for further methodological research as criteria are developed and implemented for a broader range of procedures.

Since its development more than 25 years ago, a consid-
erable amount of research has been conducted on the
reliability and validity of using the RAND/UCLA appropri-
ateness method to assess for the appropriate use of surgical procedures. Much of this research is focused on areas of controversy or concern regarding key elements of the method, such as potential variability in the process because of panelist composition, the role of the chairperson, and the necessity of having two rounds of rating separated by an in-
person discussion [8]. The purpose of our study was to summarize the results of these methodological studies. Our goal was to determine if evidence on the reliability and validity of this method supports the call for a broad and coordinated effort to develop, implement, and maintain appropriateness criteria to address variation in the use of surgical procedures.

2. Methods

2.1. RAND/UCLA appropriateness method

The RAND/UCLA appropriateness method was devel-
oped to systematically assess variation in the use of proce-
dures by defining which patients should and should not
undergo surgical intervention vs. medical therapy. An ap-
propriate indication for a procedure is one for which “the
expected health benefit (e.g., increased life expectancy, relief
of pain, reduction in anxiety, and improved functional ca-
pacity) exceeds the expected negative consequences (e.g., mor-
tality, morbidity, anxiety, pain, and time lost from work) by
a sufficiently wide margin that the procedure is worth doing”
[9]. This method starts with an extensive review of the liter-
ature on risks and benefits of the procedure. A comprehen-
sive and mutually exclusive set of clinical scenarios or
indications for the procedure is then compiled, complete with
specific definitions for any potentially ambiguous terms
(e.g., “failed medical therapy” would be explicitly defined).
Because of the need to be inclusive, the list typically includes
many hundreds of clinical indications.
Individuals comprising an expert panel rate each indica-
tion in two 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 among the panelists), or “inappropriate”
(the expected harms outweigh the expected benefits). Ap-
propriate indications are sometimes further classified as
“necessary” by the panel, usually in a third round. An in-
dication is considered necessary if it would be improper
care to not offer the procedure, there is a reasonable chance
the procedure will benefit the patient, and the magnitude of
the benefit is not small [7]. Table 1 lists examples of indi-
cations with each of these classifications [10,11].
Overuse and underuse of a surgical procedure are deter-
mained by applying the aforementioned indications to actual
patients. Underuse is defined as a 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 derived from patients who received the procedure.

2.2. Literature search, study inclusion, and data abstraction

We searched MEDLINE in February 2010 and May 2011 for articles on the appropriateness of surgical procedures using the following search strategy: MeSH terms “transplantation” OR “surgery” OR “surgical procedures, operative” AND “appropriateness” in the title or abstract. To be included, articles had to be an original research study addressing the reliability and/or validity of the RAND/UCLA appropriateness method for a surgical procedure. We also included articles that assessed the effect of altering panelist composition or eliminating the in-person discussion between rating rounds. Non-English articles were excluded. Procedures that are not commonly performed by surgeons were excluded (i.e., bronchoscopy and percutaneous coronary intervention), as were those related to pregnancy or childbirth (i.e., cesarean section and circumcision). Two physician reviewers (E.H.L. and M.M.G.) screened each study. References were mined to identify additional articles for inclusion. Disagreements were resolved by consensus.

2.2.1. Reliability of the RAND/UCLA appropriateness method

Studies on test–retest reliability were included if the same panel rated indications after allowing a period of at least 6 months time to elapse. This period was chosen based on the evidence that by 1 year, 15% of systematic reviews may be out of date [12]. Studies on interpanel reliability (i.e., the reproducibility of the RAND/UCLA appropriateness method panel results between different panels) were included if they had an experimental design. Observational studies comparing the new panel results with prior panel results were not considered to be assessing interpanel reliability because of the potential for other changing variables (panelist nationality or discipline and year study was conducted) to confound the results.

2.2.2. Construct validity of the RAND/UCLA appropriateness method

Studies on construct validity were included if they compared results produced by the RAND/UCLA appropriateness method with guideline recommendations from a professional society or with published evidence from trials or observational studies. Additionally, we included studies that compared the results of the appropriateness method against a quantitative method for predicting ratios of risks and benefits, such as decision analytic models.

2.2.3. Predictive validity of the RAND/UCLA appropriateness method

Studies on the predictive validity of the RAND/UCLA appropriateness method were included if they assessed the association between treatment concordant with appropriateness
criteria classification and clinical outcomes for patients who did and did not go on to receive a procedure. Studies that only report clinical outcomes for patients who underwent a procedure were excluded because of the potential for selection bias in these studies. Additionally, we included a study that assessed the concordance of appropriateness ratings for a procedure produced by a panel with subsequent evidence regarding appropriateness produced by randomized controlled trials because we considered this as a novel assessment of predictive validity.

2.2.4. Effect of varying panelist discipline/nationality or eliminating the in-person discussion

Studies examining the effect of varying panelist discipline and/or nationality on interpanel results were included; however, we did not review in detail the large number of studies that only compared intrapanel results by discipline or nationality; rather, these are briefly summarized regarding their findings. Experimental studies comparing the results produced by panels with and without an in-person discussion between rating rounds (i.e., in-person panels vs. panels conducted entirely by mail) were included if they otherwise followed the RAND/UCLA appropriateness method for both panels, including two rounds of rating with feedback of the first round results. As with interpanel reliability, observational studies comparing new panel results with prior panel results were excluded.

Data abstracted included dates of study and publication, procedure, study design, and findings. A review protocol was developed before initiating the search; however, it was not registered. This study received no external funding. One author (E.H.L.) was funded by the American College of Surgeons through the Robert Wood Johnson Foundation Clinical Scholars Program (RWJF CSP), and the other three authors participated through their roles as advisors to the RWJF CSP.

3. Results

3.1. Description of studies identified by the literature search

Our search identified 1,601 articles, of which 395 were screened and 37 were included in this review (Fig. 1). Three of the included articles were identified through reference mining. Some articles reported on more than one topic. Of the included articles, 5 assessed reliability, 21 assessed validity, and 17 assessed the effect of altering panel composition or eliminating the in-person discussion between rating rounds. Specifically, the test–retest reliability was reported by four articles [13−16], interpanel reliability by one article [17], comparison with guideline recommendations by nine articles [18−26], comparison with published evidence by three articles [16,27,28], comparison with a form of decision analysis by four articles [29−32], and predictive validity by four articles [20,23,33,34]. We identified one article that reported the concordance between RAND/UCLA appropriateness method panel results and subsequent randomized controlled trial results [35]. Three and 10 articles described the effect of varying panelist discipline [36−38] or nationality [24,30,39−46], respectively, and four articles compared traditional panel results with those produced by panels conducted entirely by mail [13,47−49]. Articles reporting on the development of appropriateness criteria using the RAND/UCLA appropriateness method (16 articles) or on the application of appropriateness criteria to assess for appropriate use, overuse, and underuse (27 articles) have been summarized elsewhere [50].

3.2. Reliability of the RAND/UCLA appropriateness method

The test–retest reliability of the RAND/UCLA appropriateness method has been studied for total knee replacement [13], total hip joint replacement [15], coronary artery bypass grafting (CABG) [14], and carotid endarterectomy (CEA) [16] (Table 2). In each study, the same panelists rerated

Fig. 1. Flow diagram outlining the inclusion and exclusion of studies from the literature search. Some studies addressed more than one topic and/or procedure. CABG, coronary artery bypass grafting.
a portion of the original rated indications after a time interval of 6 months to 1 year. Indications chosen for rerating were either among the most frequently found indications in clinical practice or were randomly selected, and the proportion of the original indications that were rerated ranged from 2% to 25%. Three studies reported a weighted kappa between 0.64 and 0.78, indicating good agreement between the original and subsequent ratings by the panel [51]. The fourth study reported correlation coefficients ranging from 0.75 to 0.96 between the original and subsequent ratings. No study reported the occurrence of complete discordance, in which an indication is first rated appropriate, then subsequently rated inappropriate, or vice versa.

The reproducibility of RAND/UCLA appropriateness method results between different panels with the same makeup of panelist discipline and nationality was reported by one article focused on CABG and hysterectomy [17] (Table 2). For each procedure, three parallel panels were assembled in an experimental fashion. Each panel followed the RAND/UCLA appropriateness method and rated the same indications (948 for CABG and 1,718 for hysterectomy). The three-way weighted kappa was 0.52 for CABG overuse indications, 0.83 for CABG underuse indications, and 0.51 for hysterectomy overuse indications.

### 3.3. Construct validity of the RAND/UCLA appropriateness method

We identified eight studies that classified the appropriateness of a surgical procedure for actual patients using both appropriateness criteria developed with the RAND/UCLA appropriateness method and published guidelines developed using other methods, such as consensus conferences (Table 3). The European studies focused on upper gastrointestinal (GI) endoscopy [18,19] or colonoscopy [24], whereas the US studies looked at CABG [20,21,23],

---

### Table 2. Studies on the reliability of the RAND/UCLA appropriateness method for surgical procedures

<table>
<thead>
<tr>
<th>Authors (year of publication)</th>
<th>Procedure studied</th>
<th>Methods</th>
<th>Indications rated (% rerated)</th>
<th>Weighted kappa (95% confidence interval)</th>
<th>Complete discordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escobar et al. (2003) [13]</td>
<td>Total knee replacement</td>
<td>Panel rerated indications most frequently found in the study author’s clinical practice 1 y later</td>
<td>624 (25)</td>
<td>0.78 (0.70–0.85)</td>
<td>0</td>
</tr>
<tr>
<td>Quintana et al. (2000) [15]</td>
<td>Total hip joint replacement</td>
<td>Panel rerated the most frequent appropriate, uncertain, and inappropriate indications found in a field study 1 y later</td>
<td>216 (21)</td>
<td>0.81 (0.68–0.95)</td>
<td>0</td>
</tr>
<tr>
<td>Hemingway et al. (1999) [14]</td>
<td>CABG</td>
<td>Panel rerated a random selection of the original indications 1 y later</td>
<td>84 (2)</td>
<td>0.64</td>
<td>NR</td>
</tr>
<tr>
<td>Merrick et al. (1987) [16]</td>
<td>CEA</td>
<td>Panel rerated indications 6 mo later: 66 for the clinical presentation “Multiple TIs, failure of medical treatment,” 33 randomly selected from 50 chosen as most frequently used in practice, and 33 from the remainder</td>
<td>675 (20)</td>
<td>Kappa NR; original and later repeated ratings had correlation coefficients ranging from 0.75 to 0.96</td>
<td>NR</td>
</tr>
</tbody>
</table>

**Reproducibility with different panels keeping panelist discipline and nationality constant**

<table>
<thead>
<tr>
<th>Authors (year of publication)</th>
<th>Procedure studied</th>
<th>Methods</th>
<th>Three-way weighted kappa</th>
<th>Rates of agreement among panels (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shekelle et al. (1998) [17]</td>
<td>CABG</td>
<td>Parallel three-way replication of the appropriateness panel process</td>
<td>948 (100)</td>
<td>Overuse indications</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Underuse indications</td>
</tr>
<tr>
<td>Hysterectomy</td>
<td>Parallel three-way replication of the appropriateness panel process</td>
<td>1,718 (100)</td>
<td>Overuse indications</td>
<td>0.51</td>
</tr>
</tbody>
</table>

**Abbreviations:** RAND/UCLA, RAND/University of California Los Angeles; CABG, coronary artery bypass grafting; CEA, carotid endarterectomy; NR, not reported.

Complete discordance = indication rated appropriate by one panel and inappropriate by the other.
Table 3. Construct validity: studies comparing appropriateness criteria with guideline recommendations for surgical procedures

<table>
<thead>
<tr>
<th>Authors (year of publication)</th>
<th>Procedure studied</th>
<th>Methods</th>
<th>Construct comparison</th>
<th>Percent of patients classifiable</th>
<th>Classification of patients (%)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Percent agreement&lt;sup&gt;a&lt;/sup&gt; (number of patients rated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaliszcz et al. (2006) [18]</td>
<td>Upper GI endoscopy</td>
<td>522 patients prospectively classified using both methods</td>
<td>EPAGE appropriateness criteria ANAES guidelines</td>
<td>70.7</td>
<td>63.0</td>
<td>10.7</td>
</tr>
<tr>
<td>Bersani et al. (2004) [19]</td>
<td>Upper GI endoscopy</td>
<td>2,300 patients prospectively classified using both methods</td>
<td>EPAGE appropriateness criteria ASGE guidelines</td>
<td>87.0</td>
<td>70.3</td>
<td>10.2</td>
</tr>
<tr>
<td>Epstein et al. (2003) [20]</td>
<td>CABG/PTCA</td>
<td>5,026 coronary angiography patients retrospectively classified using both methods (stratified random sample)</td>
<td>RAND appropriateness criteria ACC/AHA guidelines</td>
<td>100</td>
<td>30.4</td>
<td>NR</td>
</tr>
<tr>
<td>Leape et al. (2003) [21]</td>
<td>CABG</td>
<td>676 CABG patients retrospectively classified using both methods (stratified random sample)</td>
<td>RAND appropriateness criteria ACC/AHA guidelines</td>
<td>100</td>
<td>76</td>
<td>15</td>
</tr>
<tr>
<td>Broder et al. (2000) [22]</td>
<td>Hysterectomy</td>
<td>497 patients retrospectively classified by both methods</td>
<td>RAND appropriateness criteria ACOG guidelines</td>
<td>100</td>
<td>53.5</td>
<td>NR (n = 71)</td>
</tr>
<tr>
<td>Ziskind et al. (1999) [23]</td>
<td>CABG</td>
<td>153 patients prospectively classified using both methods</td>
<td>RAND appropriateness criteria ACC/AHA guidelines University of Maryland RAS</td>
<td>100</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>Froehlich et al. (1998) [24]</td>
<td>Colonoscopy</td>
<td>553 patients prospectively classified using both methods</td>
<td>VHS appropriateness criteria (United States) VHS/RAND appropriateness criteria (Swiss)</td>
<td>97.6</td>
<td>72.4 (appropriate or equivocal)</td>
<td>27.6</td>
</tr>
<tr>
<td>Kahn et al. (1992) [26]</td>
<td>Upper GI endoscopy</td>
<td>1,585 patients retrospectively classified using both methods (random sample)</td>
<td>ASGE guidelines RAND appropriateness criteria ASGE guidelines</td>
<td>71.6</td>
<td>72.2</td>
<td>6.7</td>
</tr>
</tbody>
</table>

Abbreviations: GI, gastrointestinal; EPAGE, European Panel on the Appropriateness of Gastrointestinal Endoscopy; ANAES, Agenee Nationale d’Accreditation et d’Evaluation en Sante (French working group); ASGE, American Society of Gastrointestinal Endoscopy; NR, not reported; CABG, coronary artery bypass grafting; PTCA, percutaneous transluminal coronary angioplasty; ACC/AHA, American College of Cardiology/American Hospital Association; ACOG, American College of Obstetrics and Gynecologists; RAS, Revascularization Appropriateness Score; VHS, Value Health Sciences.

<sup>a</sup> Only among patients able to be classified by both systems, unless specified.

<sup>b</sup> A total of 14.5% of patients were not classifiable.
hysterectomy [22], and upper GI endoscopy [26]. The number of patients studied varied from 153 to 5,026, whereas the number of patients able to be classified by both systems varied from 71 to 2,000. Of patients classified by both systems, the rates of necessary, appropriate, equivocal, and inappropriate indications were similar. For example, Leape et al. classified 676 patients who underwent CABG and reported rates of appropriate and inappropriate use of 76% and 9% by appropriateness criteria and 84% and 1.5% by American College of Cardiology/American Hospital Association guidelines. Kahn et al. classified 1,115 patients who underwent upper GI endoscopy and reported rates of appropriate and inappropriate use of 90.1% and 6.7% by appropriateness criteria and 93.5% and 3.7% by guidelines. Appropriateness criteria and guidelines thus appear to be measuring the same construct. Four of these studies found that appropriateness criteria were able to classify more patients than guidelines, whereas two European studies on upper GI endoscopy found that guidelines classified more patients, and two US studies on CABG found that all patients were classifiable by both methods.

Construct validity has also been studied by comparing RAND/UCLA appropriateness method ratings with published evidence from trials and observational studies. In a study by Merrick et al. [16], 675 indications for CEA were classified using the appropriateness method and then placed in rank order. These same indications were also ranked by the percent of recommendations for surgery found in the literature. The study authors reported that the pattern of ratings assigned by the panelists and the rank ordering were nearly identical to endorsement patterns found in the literature. Two European studies compared classification of colonoscopy indications using appropriateness criteria vs. published literature and reported weighted kappas of 0.63 (48 indications) and 0.29 (95 indications), respectively. Complete discordance occurred for 6.3% and 7.4% of indications. A comparison with US appropriateness criteria resulted in a kappa of 0.74 (47 indications). One study attempted to classify appropriateness for 577 actual patients who underwent colonoscopy and reported that only 9% of patients were classifiable by both methods [27,28].

Finally, the RAND/UCLA appropriateness method has been compared with quantitative methods for classifying indications as appropriate or inappropriate for a surgical procedure based on predictions of the ratio of risks and benefits. The results of these studies are mixed. A study comparing classification using appropriateness criteria vs. probability estimation based on the same panel’s assessment of the effect of abdominal aorta aneurysm surgery on the probability of 5-year survival reported a kappa of 0.28 [29]. Another study used a similar method for CEA and reported that the Spearman rank order correlations were significant and positive for only two of the eight panelists, with correlations ranging from 0.45 to 0.38 [32]. In contrast, a study that compared the RAND/UCLA appropriateness method with a decision model developed by asking panelists to estimate probabilities and utilities for CEA indications reported a Spearman correlation coefficient of 0.8 [31]. Finally, one study classified the appropriateness of CABG for 617 US patients and 1,053 Dutch patients using US appropriateness criteria, Dutch appropriateness criteria, and a decision analytic model built using data from randomized controlled trials and best judgment data. The authors report a kappa of 0.18 for the US appropriateness criteria vs. the model and a kappa of 0.79 for the Dutch appropriateness criteria vs. the model [30].

3.4. Predictive validity of the RAND/UCLA appropriateness method

We identified four studies on the predictive validity of the RAND/UCLA appropriateness method, all of which focused on cardiac revascularization (Table 4). All four studies were observational and examined the association between care that is concordant or discordant with appropriateness criteria classification and outcomes in patients who did and did not receive revascularization after coronary angiography. We did not identify any randomized controlled trials studying predictive validity, likely because of the ethical challenges to randomizing patients to discordant care. Two of the included studies examined mortality rates for patients classified as necessary for revascularization and found significantly lower mortality for patients who underwent revascularization compared with those who were managed medically (9% vs. 19% and 9.1% vs. 23.3%, both \( P < 0.05 \)) [20,34]. Additionally, one study reported that mortality was significantly lower for patients classified as inappropriate for CABG and did not undergo the procedure compared with those who nonetheless underwent CABG (11.9% vs. 20%, \( P < 0.05 \)) [33], whereas another study reported a non-significant trend (3% vs. 6% [20]. A large prospective study reported that patients appropriate for CABG who did not receive the procedure were more likely to have angina at follow-up (odds ratio, 3.03; 95% confidence interval, 2.08−4.42) than those who underwent revascularization and that the risk of death or nonfatal myocardial infarction at 2-year follow-up was 21% compared with 6% (\( P < 0.001 \)). Furthermore, the study authors described a graded relationship between rating and outcome over the entire scale of appropriateness (linear trend \( P = 0.002 \)) [33]. We did identify one study that found no association between concordant vs. discordant care and mortality; however, this study had a smaller sample size than the other three (153 patients vs. 2,552−5,026 patients) [23].

We identified one study that compared the results of previously developed appropriateness criteria for CEA with the results of subsequently published randomized controlled trials [35]. The results of the randomized controlled trials were concordant with the appropriateness ratings previously produced by a panel for 44 indications, which together covered almost 30% of operations performed in 1981, when the panel was conducted. Furthermore, no
indication ratings were refuted by the subsequent randomized controlled trial results.

3.5. Effect of varying panelist discipline and/or nationality

We identified two studies that compared the results of multidisciplinary panels (recommended by the RAND/UCLA appropriateness method) with those produced by all-surgeon panels. An experimental study focused on cholecystectomy found that the all-surgeon panel rated more indications appropriate (29% vs. 13%) and less inappropriate (27% vs. 50%) than the multidisciplinary panel. The study authors noted that the multidisciplinary panelists were more likely to change their ratings in the second round (after discussion of the first round results). We found three studies that compared ratings produced by the traditional RAND/UCLA appropriateness method with those produced by panels that used the appropriateness method but omitted the in-person discussion between rounds (i.e., the process was conducted entirely by mail). One experimental study compared a mail-in-only panel with three in-person panels and reported kappas of 0.49—0.67 for CABG overuse indications, 0.69—0.76 for CABG underuse indications, and 0.59—0.69 for hysterectomy overuse indications [47]. Two experimental studies focused on total knee replacement [13] and cataract extraction [48] and reported weighted kappas of 0.75 and 0.65, respectively. We also identified a study that applied the cataract extraction appropriateness criteria to 1,020 actual patients and found that the mail-in-only panel ratings classified 70% as appropriate and 3.5% inappropriate compared with 92% as appropriate and 2% as inappropriate by the in-person panel ratings (P < 0.001) [49].

Table 4. Studies on the predictive validity of appropriateness criteria for cardiac revascularization

<table>
<thead>
<tr>
<th>Authors (year of publication)</th>
<th>Patients</th>
<th>Appropriateness criteria classification</th>
<th>Number of patients with treatment concordant with appropriateness criteria classification and (% mortality)</th>
<th>Number of patients with treatment discordant with appropriateness criteria classification and (% mortality)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epstein et al. (2003) [20]</td>
<td>5,026 coronary angiography patients retrospectively classified (stratified random sample)</td>
<td>Necessary for revascularization Inappropriate for revascularization</td>
<td>n = 1,057 (9%) n = 1,425 (3%)</td>
<td>n = 469 (19%) n = 116 (6%)</td>
<td>P &lt; 0.01</td>
</tr>
<tr>
<td>Hemingway et al. (2001) [33]</td>
<td>2,552 coronary angiography patients prospectively classified</td>
<td>Appropriate for CABG Inappropriate for CABG</td>
<td>n = 765 (5.5%) n = 109 (11.9%)</td>
<td>n = 354 (19.2%) n = 15 (20%)</td>
<td>P = 0.17</td>
</tr>
<tr>
<td>Ziskind et al. (1999) [23]</td>
<td>153 coronary angiography patients prospectively classified</td>
<td>Appropriate or inappropriate for revascularization</td>
<td>n = 84 (10%)</td>
<td>n = 38 (13%)</td>
<td>P &gt; 0.05</td>
</tr>
<tr>
<td>Kravitz et al. (1995) [34]</td>
<td>4,226 coronary angiography patients retrospectively classified (random sample)</td>
<td>Necessary for CABG Necessary for revascularization</td>
<td>n = 248 (9.7%) n = 110 (9.1%)</td>
<td>n = 108 (16.7%) n = 30 (23.3%)</td>
<td>P = 0.04</td>
</tr>
</tbody>
</table>

Abbreviation: CABG, coronary artery bypass grafting. Revascularization = CABG or percutaneous transluminal coronary angioplasty; Discordant = procedure necessary or appropriate but patient treated medically or received the procedure classified as inappropriate.

Our review summarizes the methodological research on use of the RAND/UCLA appropriateness method for
surgical procedures. We found that the appropriateness method has moderate to very good reliability for determining overuse and underuse for total knee and hip joint replacement, CABG, CEA, and hysterectomy. Additionally, studies on construct validity demonstrate that appropriateness criteria developed using the RAND/UCLA appropriateness method, professional society guidelines, and evidence-based approaches are likely measuring similar constructs for overuse, underuse, and appropriate use of upper GI endoscopy, colonoscopy, CABG, hysterectomy, and CEA. However, appropriateness criteria is more often applicable to a wider range of patients than are professional society guidelines and other evidence-based approaches. Perhaps most important, though, is our finding that studies support the predictive validity of appropriateness criteria for cardiac revascularization, meaning that patients who receive treatment concordant with appropriateness criteria classification have better outcomes than those who receive discordant care. Although these findings are encouraging, they also highlight the need for further methodological research on a broader range of procedures.

The results of studies on the reliability of the RAND/UCLA appropriateness method can be put in perspective by considering appropriateness criteria in the context of a diagnostic test. This interpretation is logical because appropriateness criteria is not meant to replace a clinicians’ judgment but rather to serve as a supplement in the decision-making process, much as laboratory and radiological studies aid in the formation of diagnoses and treatment plans. Although the reliability of the appropriateness method for surgical procedures is not perfect, it is certainly within the range seen with commonly used diagnostic tests. For example, the reliability of coronary angiography for determining the presence or absence of stenosis is reported to be moderate (kappa, 0.53) [54], and the reliability for determining stenosis length and lesion eccentricity is fair (kappa, 0.38 and 0.25, respectively) [55]. The reliability of the RAND/UCLA appropriateness method panel process may also be better than that of individual surgeons’ decisions. When this was studied for hysterectomy, the reliability of individual surgeons’ decisions was fair (kappa, 0.23) [56].

Other methods of determining the appropriateness of a surgical procedure for a particular patient include guidelines produced by professional societies, evidence-based approaches using published literature, and quantitative approaches, such as decision analysis and probability estimation. There is much less published evidence on the reliability and validity of these methods. In general, studies comparing the RAND/UCLA appropriateness method with guidelines or published evidence found similar results, thus supporting the construct validity of the appropriateness method. The disadvantage of guidelines is that they tend to offer general recommendations and are thus often not actionable at the individual level. Ideally, an evidence-based approach would be taken for each patient; however, conducting randomized controlled trials for every possible clinical scenario is not realistic and may not even be ethical in some circumstances. In contrast, the RAND/UCLA appropriateness method is designed to be comprehensive and exhaustive, producing appropriateness criteria that can be applied to the vast majority of patients presenting for a procedure. Decision models and probability estimation may be similarly comprehensive and actionable. As with the appropriateness method, these methods rely on a synthesis of published evidence and clinical judgment. We found relatively few studies comparing the RAND/UCLA appropriateness method with these methods and believe that this is an area in need of further study.

There are a number of ways in which appropriateness criteria could be implemented to reduce variation in the use of surgical procedures. For example, in response to allegations of inappropriate use of percutaneous coronary intervention in Maryland hospitals, the American College of Cardiology and the Society for Cardiovascular Angiography and Interventions recently proposed a mandatory accrediting process through the Accreditation for Cardiovascular Excellence (ACE) program. Along with various process and outcome standards, a key component of the ACE program is documentation of the indication for percutaneous coronary intervention and assessment of appropriateness using the current Appropriate Use Criteria for Coronary Artery Revascularization [57,58]. Appropriateness criteria could also be used by physicians for clinical decision support or implemented in the clinical realm as part of the preoperative informed consent process, with patients receiving individualized assessments of the appropriateness of the procedure for their particular clinical scenario. For example, a proof of concept of the utility of appropriateness criteria for changing physician behavior and reducing practice variations was performed in the United Kingdom using criteria for coronary angiography [59]. These uses would be consistent with recent quality measures proposed by the Centers for Medicare and Medicaid Services for Accountable Care Organizations, which include a measure of the percentage of physicians using clinical decision support and a measure of shared decision-making between physician and patient.

Our study has possible limitations. First, methodological studies on the RAND/UCLA appropriateness method for surgical procedures may have been overlooked. To reduce this possibility, we started with broad search terms and had two physicians to perform the screening and reference mining. Second, the implications of our findings may be limited by the possibility that the RAND/UCLA appropriateness method may have differing reliability and validity for different procedures. Most of the methodological studies on the RAND/UCLA appropriateness method focus on a relatively small number of procedures. Further evaluation of the method is warranted and could be performed concurrently with future development, application, and implementation of appropriateness criteria for a broad range of procedures. Finally, the developers of the RAND/UCLA
appropriateness method performed many of the methodological studies we identified, which may be a source of bias. Of note, however, a key study on the predictive validity of the RAND/UCLA appropriateness method for coronary revascularization was performed completely independent of the appropriateness method developers [33]. Ensuring that patients receive appropriate surgical care should be considered integral to improving the overall quality of the health care system. Our study supports the use of the RAND/UCLA appropriateness method to assess variation in the use of surgical procedures by identifying overuse and underuse and highlights the need for further methodological research as appropriateness criteria is developed and implemented for a broader range of procedures.

Acknowledgments

The authors acknowledge Dr Angela Ingraham for her help with the initial screening of articles. This study received no external funding. Dr Lawson’s time was supported by the American College of Surgeons through RWJF CSP. The remaining three authors participated through their roles as advisors to the RWJF CSP.

References
