

Special Articles

THE REPRODUCIBILITY OF A METHOD TO IDENTIFY THE OVERUSE AND UNDERUSE OF MEDICAL PROCEDURES

PAUL G. SHEKELLE, M.D., PH.D., JAMES P. KAHAN, PH.D., STEVEN J. BERNSTEIN, M.D., M.P.H., LUCIAN L. LEAPE, M.D., CAREN J. KAMBERG, M.P.H., AND R.E. PARK, PH.D.

ABSTRACT

Background To assess the overuse and underuse of medical procedures, various methods have been developed, but their reproducibility has not been evaluated. This study estimates the reproducibility of one commonly used method.

Methods We performed a parallel, three-way replication of the RAND–University of California at Los Angeles appropriateness method as applied to two medical procedures, coronary revascularization and hysterectomy. Three nine-member multidisciplinary panels of experts were composed for each procedure by stratified random sampling from a list of experts nominated by the relevant specialty societies. Each panel independently rated the same set of clinical scenarios in terms of the appropriateness of the relevant procedure on a risk–benefit scale ranging from 1 to 9. Final ratings were used to classify the procedure in each scenario as necessary or not necessary (to evaluate underuse) and inappropriate or not inappropriate (to evaluate overuse). Reproducibility was measured by overall agreement and by the kappa statistic. The criteria for underuse and overuse derived from these ratings were then applied to real populations of patients who had undergone coronary revascularization or hysterectomy.

Results The rates of agreement among the three coronary-revascularization panels were 95, 94, and 96 percent for inappropriate-use scenarios and 93, 92, and 92 percent for necessary-use scenarios. Agreement among the three hysterectomy panels was 88, 70, and 74 percent for inappropriate-use scenarios. Scenarios involving necessary use of hysterectomy were not assessed. The three-way kappa statistic to detect overuse was 0.52 for coronary revascularization and 0.51 for hysterectomy. The three-way kappa statistic to detect underuse of coronary revascularization was 0.83. Application of individual panels' criteria to real populations of patients resulted in a 100 percent variation in the proportion of cases classified as inappropriate and a 20 percent variation in the proportion of cases classified as necessary.

Conclusions The appropriateness method is far from perfect. Appropriateness criteria may be useful in comparing levels of appropriate procedures among populations but should not by themselves be used to direct care for individual patients. (N Engl J Med 1998; 338:1888-95.)

©1998, Massachusetts Medical Society.

THE appropriateness of health procedures has commanded considerable attention recently.¹⁻³ Escalating health care costs and identification of inappropriate care have led to the critical examination of possible overuse and underuse of many medical and surgical procedures and questions as to when or whether they are needed. Central to this examination is the determination of what constitutes appropriate indications for any given procedure. Ideally, this determination would be derived solely from rigorously conducted research that established conclusively the clinical circumstances under which patients benefit from the procedure. Unfortunately, satisfactory data on efficacy and effectiveness are unusual.⁴ In fact, several studies estimate that only 15 to 20 percent of medical practices can be justified on the basis of rigorous scientific data establishing their effectiveness.^{5,6} For most conditions, something other than rigorous data on efficacy or effectiveness must be used to determine criteria of appropriateness.

One frequently used method that combines expert opinion, the type of information most commonly employed, with available scientific evidence is the RAND–University of California at Los Angeles appropriateness method, which was developed in 1984 by the Health Services Utilization Study.⁷ This method has been used to evaluate the appropriateness of a variety of medical and surgical interventions.^{1,3,8,9} It combines a systematic review of the scientific literature with expert opinion and yields specific criteria of appropriateness that can be used as the basis for review criteria, practice guidelines, or both. In general, it quantitatively assesses the expert judgment of a multidisciplinary group of clinicians concerning a comprehensive series of clinical indications on a risk–benefit scale ranging from 1 to 9. It

From the West Los Angeles Veterans Affairs Medical Center, Los Angeles (P.G.S.); RAND, Santa Monica, Calif. (P.G.S., J.P.K., C.J.K., R.E.P.); the Ann Arbor Veterans Affairs Medical Center and the Departments of Internal Medicine and Health Management and Policy, University of Michigan, Ann Arbor (S.J.B.); and the Harvard School of Public Health, Boston (L.L.L.). Address reprint requests to Dr. Shekelle at RAND, 1700 Main St., P.O. Box 2138, Santa Monica, CA 90407-2138.

is iterative, with two rounds of anonymous ratings and a face-to-face group discussion between rounds. Each panelist has equal weight in determining the final result: an explicit appropriateness rating for clinically detailed patient scenarios.

A central criticism of the appropriateness method is the potential sensitivity of the results to the selection of particular experts, leading to concern about the results' validity.^{10,11} To address this concern, we conducted a rigorous test of the reproducibility of the appropriateness method as used to identify the overuse and underuse of medical procedures.

METHODS

We performed a parallel, three-way replication of the appropriateness panel process for two medical procedures, coronary revascularization and hysterectomy. We chose these procedures because they are commonly performed and they differ in the amount of available scientific evidence concerning efficacy. We examined all indications for coronary revascularization (948 clinical scenarios) and nonemergency, nononcologic indications for hysterectomy (1718 clinical scenarios). Table 1 presents examples of indications that were rated.

Selection of Panelists

We solicited nominations for the coronary-revascularization and hysterectomy panels from a variety of relevant, respected medical and surgical societies and organizations. From all sources, 69 cardiologists, 30 primary care physicians, and 81 cardiovascular surgeons were nominated for the coronary-revascularization panel, and 57 obstetrician-gynecologists and 30 primary care physicians were nominated for the hysterectomy panel.

We requested a current curriculum vitae from each nominee. Physicians who had previously served as expert panelists for assessments of the appropriateness of coronary revascularization or hysterectomy were excluded. Each panelist was classified according to specialty, location of practice, type of practice (academic or private), and sex. Drawing from the pool of qualified nominees by stratified random sampling, we made assignments to four panels for each procedure. We sent the panelists who were selected a letter inviting them to participate. Those who declined were replaced with new physicians from the appropriate strata until four panels for each procedure had been composed. Our interaction with one of these panels was only by mail. We report here the results from the three panels that followed the conventional appropriateness method, which includes a face-to-face panel discussion.

Synthesis of the Literature and Selection of Moderators

For each procedure, a synthesis of the scientific literature was prepared and peer-reviewed by external experts for completeness and accuracy. Three experienced moderators were selected, one for each panel. Moderators were aware only of the names of their own panelists and their own results; they were unaware of the names of other panelists and of the actions and results of the other panels.

Operation of the Panels

Each panel was conducted in identical fashion, with panelists receiving the same literature synthesis, set of clinical scenarios, and instructions. The panelists first independently rated the appropriateness of using the relevant procedure in each scenario and returned their rating forms by mail. The ratings were then tabulated before the face-to-face panel meeting. Each coronary-revascularization panel had a 2-day face-to-face meeting (all three of which took place over a 10-day span in October 1994). Likewise, the three hysterectomy panels met independently for two days

TABLE 1. EXAMPLES OF THE INDICATIONS FOR CORONARY REVASCULARIZATION AND HYSTERECTOMY RATED BY EXPERT PANELS.*

Coronary revascularization

1. A patient with severe angina (class III or IV) receiving maximal medical therapy, with two-vessel disease without involvement of the proximal left anterior descending artery, who has strongly positive results on exercise electrocardiography, has an ejection fraction of 20 to 35%, and is at low risk with regard to PTCA and surgery.
2. An asymptomatic patient, within 21 days after an acute transmural (Q-wave) myocardial infarction, with single-vessel disease of the proximal left anterior descending artery, who has less than strongly positive results on or did not undergo exercise electrocardiography, has an ejection fraction of >35%, and is at high risk with regard to PTCA and surgery.
3. An asymptomatic patient with three-vessel disease who has an ejection fraction of >50% and is at low risk with regard to PTCA and surgery.

Hysterectomy

1. A patient with grade I or II cervical dysplasia, 40 years of age or older, who has undergone one conization or excision with margins of resection showing dysplasia, in whom further sampling shows persistent dysplasia.
2. A patient who currently has abnormal uterine bleeding of unknown cause, is less than 40 years old, has undergone an endometrial biopsy, has received hormone treatment, and has a mild degree of anemia with major functional impairment.
3. A premenopausal patient with leiomyomas, bleeding, and pain or discomfort, who is 40 or older, has a uterine size corresponding to more than 14 weeks' gestation, is currently bleeding, has had an endometrial biopsy, received medical treatment for pain or discomfort or both, and has had neither clinically significant anemia nor major functional impairment.

*Panelists were given explicit definitions of all terms such as "strongly positive results," "high risk with regard to surgery," and "medical treatment for pain or discomfort." PTCA denotes percutaneous transluminal coronary angiography.

each in November 1994. All panel meetings occurred in the same room at the RAND office in Washington, D.C. In the only departure from usual practice, we did not allow panelists to alter clinical scenarios, because we wanted an identical set of scenarios in order to compare results among panels. To minimize the potential effect of this change, we extensively tested our scenarios with nonpanelists for clinical sensibility before we used them.

After obtaining the final-round appropriateness ratings, we had the coronary-revascularization panelists rate again each scenario that they had judged appropriate for use of the relevant procedure, this time according to necessity criteria. The concept of necessity goes beyond that of appropriateness, in that withholding a procedure that was deemed necessary for a person's clinical situation would constitute wrongful underuse of the procedure.¹² Because our study was restricted to the use of hysterectomy for nonemergency, nononcologic indications, we did not ask the hysterectomy panel for necessity ratings.

Statistical Analysis

With final ratings from each panel, we assigned an appropriateness category to each clinical indication. Disagreement was considered to have occurred when at least three panelists rated an indication in the top third of the risk-benefit scale (7, 8, or 9) and at least three panelists rated the same indication in the bottom third (1, 2, or 3). A median panel rating of 7, 8, or 9 without disagreement defined an indication as appropriate. A median panel rating of 1, 2, or 3 without disagreement defined an indication as inappropriate. Indications with a median rating of 4, 5, or 6, and all indications with disagreement, were classified as uncertain. Indications judged appropriate with a median panel rating of 7, 8, or 9 on the necessity scale without disagreement were considered evidence of a procedure's necessity.

TABLE 2. COMPOSITION OF THE EXPERT PANELS.

CHARACTERISTICS OF THE PANELISTS	PANEL	PANEL	PANEL
	A	B	C
	number of panelists		
Coronary revascularization			
Type of physician*			
Cardiovascular surgeon	3	3	3
PTCA-performing cardiologist	3	3	3
Non-PTCA-performing cardiologist	1	1	1
Primary care physician	2	2	2
Type of practice			
Private	3	3	3
Academic	6	6	6
Sex			
Female	1	1	1
Male	8	8	8
Geographic location			
East	2	3	2
Midwest	3	2	3
South	2	2	2
West	2	2	2
Hysterectomy			
Type of physician			
Operating gynecologist	4	4	4
Nonoperating gynecologist	2	2	2
Primary care physician	3	3	3
Type of practice			
Private	3	3	3
Academic	6	6	6
Sex			
Female	3	3	2
Male	6	6	7
Geographic location			
East	3	3	2
Midwest	2	2	2
South	2	2	3
West	2	2	2

*PTCA denotes percutaneous transluminal coronary angioplasty.

We analyzed the final-round appropriateness ratings using the pairwise percentage of agreement between panels, the kappa statistic (a measure of agreement that takes into account the agreement due to chance), and the three-way kappa statistic among panels. We used terminology suggested by Landis and Koch¹³ to assign descriptive terms to numerical values of kappa. To identify overuse, we used the ratings to classify each procedure as “inappropriate” or “not inappropriate.” To identify underuse of coronary revascularization, we used the classification of “necessary” or “not necessary.” These classifications are the same as those used in previous studies of overuse and underuse. For each calculation, the indication was weighted by the frequency with which it occurs in practice. For the weights for overuse of coronary revascularization, we used data from 2532 persons (randomly selected from 15 hospitals in New York State) who had undergone coronary revascularization. For the weights for underuse of coronary revascularization, we used data from 1294 persons (randomly selected from 15 New York hospitals) who had undergone coronary angiography. For hysterectomy, we used data from 636 women (randomly selected from seven managed-care organizations) who had undergone hysterectomy for nonemergency, nononcologic indications. The methods used for collecting data and assigning appropriateness criteria based on medical records have been previously reported.¹⁻³ In brief, clinical data were collected from the medical records in sufficient detail to allow each case to be matched with one of the clinical scenarios rated by the panels for appropriateness.

Stata software (version 5.0, Stata, College Station, Tex.) was used for calculations. Confidence intervals were calculated by the bias-corrected bootstrap method.

RESULTS

Participation rates were extremely high among those invited to serve as panel members. Of the cardiovascular panelists invited, 98 percent agreed to participate, and of the hysterectomy panelists, 91 percent agreed to participate. The three panels for each procedure were well matched with regard to all measured characteristics (Table 2).

The respective final-round ratings of Panels A, B, and C showed disagreement on 1, 4, and 4 percent of the coronary-revascularization scenarios and 9, 6, and 2 percent of the hysterectomy scenarios.

The degree of agreement on appropriateness among the panels was mixed. Table 3 shows the pairwise agreement, pairwise kappa statistic, and three-way kappa statistic for overuse and underuse. For coronary revascularization, there were high levels of agreement among panels, with moderate agreement beyond chance with regard to overuse and almost perfect agreement beyond chance with regard to underuse. For hysterectomy, Panels A and B had a very high level of agreement, and substantial agreement beyond chance, with regard to overuse. Panel C had a lower level of overall agreement with the other two panels. For both procedures, the three-way agreement beyond chance with regard to overuse was moderate, and for coronary revascularization, the three-way agreement beyond chance with regard to underuse was almost perfect.

Figure 1 shows the effect of using the appropriateness ratings of the three coronary-revascularization panels to classify the 2532 cases of coronary revascularization in New York. Had Panel A’s ratings alone been used to classify care, 160 procedures would have been labeled as inappropriate. Of these, none would have been rated as necessary by either of the other two panels, and 18 would have been rated as appropriate by one of the other panels. Similarly, if Panel B’s ratings alone had been used to classify care, 186 procedures would have been labeled as inappropriate, and none of these would have been rated as necessary or appropriate by either of the other two panels. Finally, if Panel C’s ratings alone had been used to classify care, 97 procedures would have been labeled as inappropriate; none of these would have been rated as necessary, but 2 would have been rated as appropriate by one of the other panels. In no instance was a case rated as necessary by one panel and inappropriate by another.

Figure 2 provides similar data about the underuse of coronary revascularization. Of 1294 uses of angiography, 498, 464, and 402 would have been rated as necessary by Panels A, B, and C, respectively. No use of angiography judged necessary by one panel

TABLE 3. COMPARISONS OF PANEL RATINGS OF OVERUSE AND UNDERUSE.*

COMPARISON	CORONARY REVASCULARIZATION		HYSTERECTOMY	
	PERCENT AGREEMENT	KAPPA STATISTIC (95% CI)	PERCENT AGREEMENT	KAPPA STATISTIC (95% CI)
Overuse				
Panels A and B	95	0.60 (0.32–0.85)	88	0.71 (0.60–0.80)
Panels B and C	94	0.40 (0.13–0.69)	70	0.41 (0.30–0.55)
Panels A and C	96	0.56 (0.25–0.82)	74	0.49 (0.35–0.63)
Three-way		0.52 (0.31–0.72)		0.51 (0.39–0.62)
Underuse				
Panels A and B	93	0.85 (0.69–0.95)		
Panels B and C	92	0.81 (0.62–0.93)		
Panels A and C	92	0.82 (0.64–0.94)		
Three-way		0.83 (0.67–0.93)		

*CI denotes confidence interval. There was no analysis of underuse of hysterectomy.

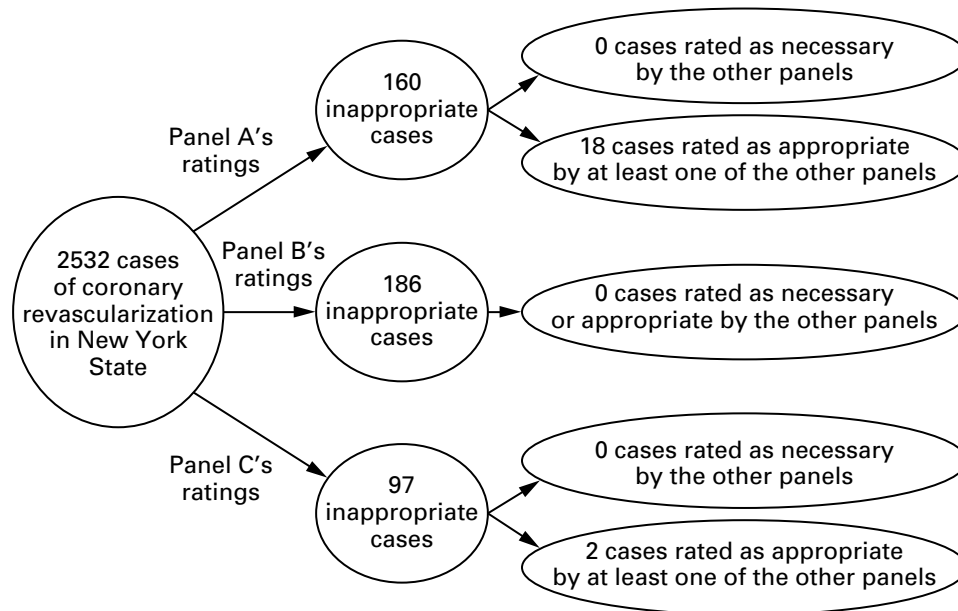


Figure 1. Effect of the Three Panels' Appropriateness Ratings on the Determination of Overuse of Coronary Revascularization.

was rated as inappropriate by either of the other two panels; some were rated as uncertain by at least one other panel (24, 31, and 4 by Panels A, B, and C, respectively).

Finally, Figure 3 shows the effect of using the appropriateness ratings of the three hysterectomy panels to classify 636 cases of hysterectomy. Using Panel A's ratings or Panel B's ratings alone would have labeled 200 or 153 hysterectomies, respectively, as inappropriate, with 7 of them for each panel rated as appropriate by one of the other two panels. Using Panel C's ratings alone would have labeled 331 hys-

terectomies as inappropriate, with 92 of them rated as appropriate by one of the other two panels.

We examined the indications for which results were discordant among panels and found none in which conclusive evidence from randomized, clinical trials supported a given action. For overuse of revascularization, three indications involved discordant ratings (in a total of 20 cases). Sixteen cases were accounted for by one indication (patients with chronic stable angina, mild or moderate angina, and single-vessel disease who had less than strongly positive results on an exercise stress test or in whom the stress

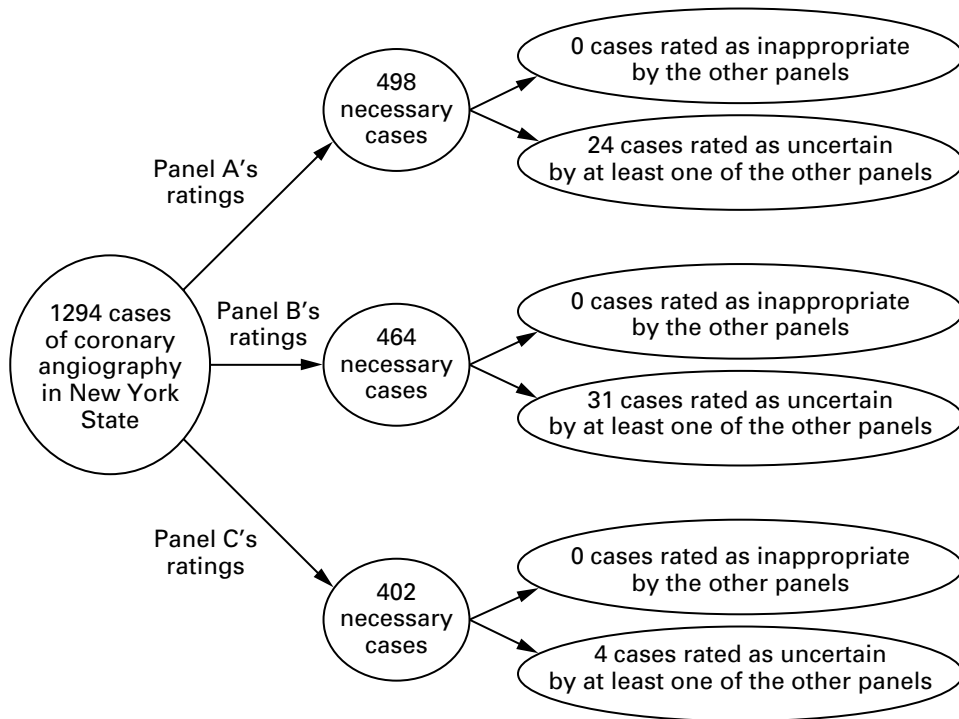


Figure 2. Effect of the Three Panels' Appropriateness Ratings on the Determination of Underuse of Coronary Revascularization.

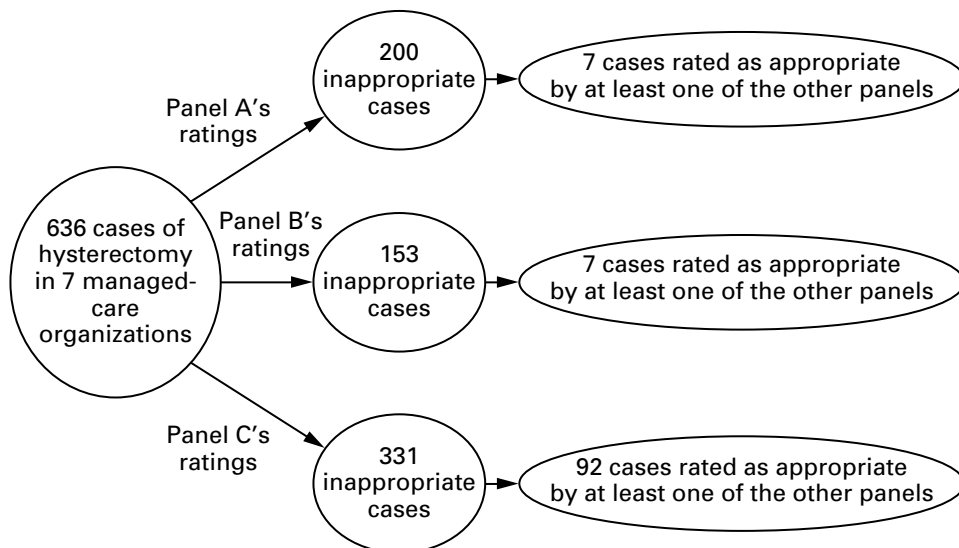


Figure 3. Effect of the Three Panels' Appropriateness Ratings on the Determination of Overuse of Hysterectomy.

test was not done). For underuse of revascularization, 13 indications involved discordant ratings (in a total of 46 cases). Four indications accounted for 35 cases, including three that involved patients presenting within 21 days after an acute myocardial infarction and one that involved asymptomatic patients with three-vessel disease. The 92 cases of hysterectomy with discordant results were spread over 28 indications, of which 26 (93 percent, involving 90 [98 percent] of the cases) involved uterine bleeding (or pelvic discomfort) with "major impairment" of the patient, which was defined as follows: "during the last 3 months the patient had had a significant worsening in level of activity (e.g., 2 or more days per month) due to her bleeding or pain, or the bleeding or pain is continuing to have a significant negative effect on her functional ability."

DISCUSSION

Our results show that the appropriateness method of identifying overuse is far from perfect. The degree of agreement among panels about care identified as inappropriate was only moderate. Furthermore, the number of cases categorized as inappropriate varied by a factor of about two for both procedures. However, our results for identifying underuse are more reassuring. Agreement among panels was nearly perfect, and the number of cases classified as necessary varied by only 20 percent among panels.

The literature is sparse on studies evaluating the reproducibility and reliability of alternative methods for determining appropriateness. We do know, however, that alternative methods are certain to be less than perfect. The reliability of individual surgeons' decisions to recommend hysterectomy has been estimated to have a kappa of 0.23.¹⁴ Although imperfect, the reproducibility of the appropriateness method is markedly better. Three-to-fivefold variations in the rate of use of hysterectomy have been reported¹⁵⁻¹⁷ and have been attributed to variability among physicians.¹⁸ Although imperfect, the appropriateness method is less variable. A recent report on coronary angiography after myocardial infarction reported a 2.5-fold variation in the rate of use among 16 Kaiser Permanente hospitals.¹⁹ For cases in which coronary angiography was judged necessary (by a process identical to that described here), there was a 1.6-fold variation. Again, although imperfect, the results of the appropriateness method for coronary revascularization are less variable.

Although systematic data are lacking, the results of other methods, such as meta-analysis, decision analysis, and cost-effectiveness analysis, have also been variable. For example, meta-analyses on the same topics have reached different conclusions,²⁰ and meta-analyses do not always agree with subsequent clinical trials.^{21,22} A recent systematic evaluation of the agreement between meta-analyses and subsequent large

clinical trials reported a kappa of 0.3.²³ Likewise, three independent decision analyses on the use of isoniazid prophylaxis for patients with positive results on tuberculin skin tests came to three different conclusions.²⁴ The estimates of the cost effectiveness of autologous blood donation have also varied greatly, even for the same surgical procedure.²⁵⁻³⁰ Whether any of these methods is more or less reliable than the appropriateness method remains to be studied systematically.

The area of medicine with the largest amount of rigorous data on reliability is diagnostic testing. Although not a diagnostic test, the appropriateness method shares many characteristics with diagnostic tests, in that both involve classifying patients into two or more categories and both therefore have a reproducibility, false positive, and false negative rate. In ischemic cardiac disease and in women's health, the reliability of thallium scintigraphy for the diagnosis of ischemic cardiac disease has been estimated to have a kappa of 0.45³¹ and a kappa of 0.66,³² the reliability of coronary angiography in determining the presence or absence of stenosis has been estimated to have a kappa of 0.53,³³ the reliability of screening mammography has been estimated to have a kappa of 0.47,³⁴ and the reliability of the classification of cervical smears with grade III histologic features has been estimated to have a kappa of 0.50³⁵ and a kappa of 0.58.³⁶ Given these values, the reproducibility of the appropriateness method is about the same as that of several well-accepted diagnostic tests.

However, the variability we observed in the appropriateness method does have important implications for clinical use. When the method is used to measure rates in a single population, the fact that the classification of inappropriate use varies by a factor of two means that precise estimates are not possible. At best, in a single population, the appropriateness method can estimate whether the proportion of cases with overuse is small or large. The appropriateness method will perform more acceptably as a way to assess the relative proportions of overuse and underuse among populations. Bias due to misclassification will be present in all comparison groups. Although the absolute measure of overuse and underuse may be biased because of misclassification, the relative difference among groups is less likely to be biased.

In making decisions for individual patients, however, the situation is different. Like diagnostic tests, the appropriateness method does not have sufficient reproducibility to justify its use as a gold standard of appropriateness. Clinicians and patients may wish to use results of the appropriateness method as a starting point for discussions about the expected net outcome of a medical procedure. Purchasers, however, should consider the appropriateness method as no more than a screening test to identify care that may

be inappropriately under- or overdelivered. Care that is so identified should then be examined at the next level, which must involve direct contact with the provider, and possibly the patient as well, to ascertain additional details about the care delivered. Under no circumstances should the care of individual patients be guided solely by the results of the appropriateness method without additional clinical information.

Our data certainly make it clear that the reproducibility of the appropriateness method could be improved. Although our results for coronary revascularization may be acceptable, we need to know whether the difference between groups of experts considering other procedures is likely to be of a magnitude similar to that seen for hysterectomy between Panel C and the other two panels. The variability in the effect of "major impairment" of function on the appropriateness ratings reflects the different way that Panel C interpreted the trade-off between risk and benefit for these patients; the symptom of major impairment was not judged sufficient to outweigh the risk of the procedure. This finding underscores the variability of physicians' interpretations of the importance of patients' symptoms (as opposed, for example, to mortality or the probability of a myocardial infarction). It also highlights the need for clinical trials of hysterectomy that directly measure symptoms as a primary outcome and the need to involve patients in quality-of-life decisions.

Further research is needed to identify which procedures are likely to be associated with reliable appropriateness-method results. We can conjecture that the more firm evidentiary basis underlying the indications for revascularization resulted in a more reliable extrapolation beyond the evidence on the part of the experts. For hysterectomy, where the evidence was scant and the judgments were dependent on individual values, reliability was reduced. This hypothesis can be further explored by examining in detail the panel discussions or analyzing the results of different panels for different procedures. Multiple determinations have also been suggested as a way to improve the reliability of some diagnostic tests, such as mammography and coronary angiography.^{33,34}

The use of stratified random sampling, the high participation rate achieved, the similarity of the panelists in many features, and the identical nature of the process in each panel all strengthen this study as a fair and rigorous test of the reproducibility of the expert-panel component of the appropriateness method. However, our study has several limitations. Additional components not tested include the development of the systematic review and the construction of the clinical scenarios, each of which may contribute to variability. Also, we studied only two procedures. Although this was a deliberate choice de-

signed to identify likely upper and lower boundaries of reproducibility (with coronary revascularization and hysterectomy, respectively), values for other procedures may be below the values reported here for hysterectomy.

Future studies of the reproducibility of methods identifying overuse and underuse of health procedures should be conducted as rigorously as the study reported here. Only then can we inform with empirical evidence what has thus far been a debate based largely on theory and opinion about how best to determine what care is appropriate.

Supported by a grant (HSO7185-02) from the Agency for Health Care Policy and Research. Dr. Shekelle is the recipient of a Senior Research Associate Career Development Award from the Department of Veterans Affairs.

We are indebted to Mark Chassin, M.D., for helpful comments and to the physicians who served as panelists.

REFERENCES

1. Leape LL, Hilborne LH, Park RE, et al. The appropriateness of use of coronary artery bypass graft surgery in New York State. *JAMA* 1993;269:753-60.
2. Bernstein SJ, Hilborne LH, Leape LL, et al. The appropriateness of use of coronary angiography in New York State. *JAMA* 1993;269:766-9.
3. Bernstein SJ, McGlynn EA, Siu AL, et al. The appropriateness of hysterectomy: a comparison of care in seven health plans. *JAMA* 1993;269:2398-402.
4. Fink A, Brook RH, Koseoff J, Chassin MR, Solomon DH. Sufficiency of clinical literature on the appropriate uses of six medical and surgical procedures. *West J Med* 1987;147:609-14.
5. Institute of Medicine. *Assessing medical technologies*. Washington, D.C.: National Academy Press, 1985.
6. Dubinsky M, Ferguson JH. Analysis of the National Institutes of Health Medicare coverage assessment. *Int J Technol Assess Health Care* 1990;6:480-8.
7. Brook RH, Chassin MR, Fink A, Solomon DH, Koseoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care* 1986;2:53-63.
8. Gray D, Hampton JR, Bernstein SJ, Koseoff J, Brook RH. Audit of coronary angiography and bypass surgery. *Lancet* 1990;335:1317-20.
9. Bengtson A, Herlitz J, Karlsson T, Brandrup-Wognsen G, Hjalmarson A. The appropriateness of performing coronary angiography and coronary artery revascularization in a Swedish population. *JAMA* 1994;271:1260-5.
10. Phelps CE. The methodologic foundations of studies of the appropriateness of medical care. *N Engl J Med* 1993;329:1241-5.
11. Hicks NR. Some observations on attempts to measure appropriateness of care. *BMJ* 1994;309:730-3.
12. Kahan JP, Bernstein SJ, Leape LL, et al. Measuring the necessity of medical procedures. *Med Care* 1994;32:357-65.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
14. Rutkow IM, Gittelsohn AM, Zuidema GD. Surgical decision making: the reliability of clinical judgment. *Ann Surg* 1979;190:409-19.
15. Roos NP. Hysterectomy: variations in rates across small areas and across physicians' practices. *Am J Public Health* 1984;74:327-35.
16. *Hysterectomies in New York State: a statistical profile*. Albany: New York State Department of Health, 1988:1-13.
17. Haas S, Acker D, Donahue C, Katz ME. Variation in hysterectomy rates across small geographic areas of Massachusetts. *Am J Obstet Gynecol* 1993;169:150-4.
18. Carlson KJ, Nichols DH, Schiff I. Indications for hysterectomy. *N Engl J Med* 1993;328:856-60.
19. Selby JV, Fireman BH, Lundstrom RJ, et al. Variation among hospitals in coronary-angiography practices and outcomes after myocardial infarction in a large health maintenance organization. *N Engl J Med* 1996;335:1888-96.
20. Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. II. Replicate vari-

ability and comparison of studies that agree and disagree. *Stat Med* 1987; 6:733-44.

21. Borzak S, Ridker PM. Discordance between meta-analyses and large-scale randomized, controlled trials: examples from the management of acute myocardial infarction. *Ann Intern Med* 1995;123:873-7.

22. Cappelleri JC, Ioannidis JP, Schmid CH, et al. Large trials vs. meta-analysis of smaller trials: how do their results compare? *JAMA* 1996;276:1332-8.

23. LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian E. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N Engl J Med* 1997;337:536-42.

24. Colice GL. Decision analysis, public health policy, and isoniazid chemoprophylaxis for young adult tuberculin skin reactors. *Arch Intern Med* 1990;150:2517-22.

25. Birkmeyer JD, AuBuchon JP, Littenberg B, et al. Cost-effectiveness of preoperative autologous donation in coronary artery bypass grafting. *Ann Thorac Surg* 1994;57:161-8.

26. Birkmeyer JD, Goodnough LT, AuBuchon JP, Noordsij PG, Littenberg B. The cost-effectiveness of preoperative autologous blood donation for total hip and knee replacement. *Transfusion* 1993;33:544-51.

27. Goodnough LT, Grishaber JE, Birkmeyer JD, Monk TG, Catalona WJ. Efficacy and cost-effectiveness of autologous blood predeposit in patients undergoing radical prostatectomy procedures. *Urology* 1994;44:226-31.

28. Kattan MW, Eastham JA, Yawn DH, Scardino PT. A decision analysis of the cost effectiveness of preoperative autologous blood donation prior

to radical prostatectomy for clinically localized prostate cancer. *Med Decis Making* 1995;15:429. abstract.

29. Etchason J, Petz L, Keeler E, et al. The cost effectiveness of preoperative autologous blood donations. *N Engl J Med* 1995;332:719-24.

30. Sonnenberg FA, Nizam RA, Yomtovian RA, et al. Cost-effectiveness of autologous blood donation revisited: the impact of increased risk of bacterial infection following allogeneic transfusion. *Med Decis Making* 1995; 15:428. abstract.

31. Wackers FJ, Bodenheimer M, Fleiss JL, Brown M. Factors affecting uniformity in interpretation of planar thallium-201 imaging in a multi-center trial. *J Am Coll Cardiol* 1993;21:1064-74.

32. Atwood JE, Jensen D, Froelicher V, et al. Agreement in human interpretation of analog thallium myocardial perfusion images. *Circulation* 1981;64:601-9.

33. DeRouen TA, Murray JA, Owen W. Variability in the analysis of coronary arteriograms. *Circulation* 1977;55:324-8.

34. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994; 331:1493-9.

35. Kato K, Santamaria M, De Ruiz PA, et al. Inter-observer variation in cytological and histological diagnoses of cervical neoplasia and its epidemiologic implication. *J Clin Epidemiol* 1995;48:1167-74.

36. Ismail SM, Colclough AB, Dinnen JS, et al. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *BMJ* 1989;298:707-10.

RECEIVE THE *JOURNAL'S* TABLE OF CONTENTS EACH WEEK BY E-MAIL

To receive the table of contents of the *New England Journal of Medicine* by e-mail every Thursday morning, send an e-mail message to:

listserv@massmed.org

Leave the subject line blank, and type the following as the body of your message:

subscribe TOC-L

You can also sign up through our website at: <http://www.nejm.org>
