

A RAND NOTE

**A Method for the Detailed Assessment of the
Appropriateness of Medical Technologies**

**Robert H. Brook, Mark R. Chassin, Arlene Fink,
David H. Solomon, Jacqueline Kosecoff,
R. E. Park**

RAND

The research described in this report was supported by a grant from the U.S. Department of Health and Human Services.

This Note contains an offprint of RAND research originally published in a journal or book. The text is reproduced here, with permission of the original publisher.

The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of RAND do not necessarily reflect the opinions or policies of the sponsors of RAND research.

A RAND NOTE

N-3376-HHS

**A Method for the Detailed Assessment of the
Appropriateness of Medical Technologies**

**Robert H. Brook, Mark R. Chassin, Arlene Fink,
David H. Solomon, Jacqueline Kosecoff,
R. E. Park**

**Supported by the
U.S. Department of Health and Human Services**

RAND

A METHOD FOR THE DETAILED ASSESSMENT OF THE APPROPRIATENESS OF MEDICAL TECHNOLOGIES

Robert H. Brook

Mark R. Chassin

Arlene Fink

David H. Solomon

University of California, Los Angeles

Rand Corporation

Jacqueline Kosecoff

R. E. Park

University of California, Los Angeles

INTRODUCTION

The standard way to assess medical technologies is to conduct a randomized clinical trial. Patients are randomly assigned to groups receiving alternative treatments, and outcomes are monitored over a long period of time. For example, some victims of left main coronary artery disease may undergo coronary artery bypass surgery, and others may receive medical treatment with nitroglycerine and beta blockers. Comparison of five-year mortality and morbidity in the two groups helps to determine the relative appropriateness of the two procedures. In addition, information about quality of life and cost can also be collected and compared.

Acknowledgments: This study was prepared under grants from the Commonwealth Fund, the Hartford Foundation, the Pew Memorial Trust, and the Robert Wood Johnson Foundation. We are most grateful for the assistance of Katherine Kahn, M.D. and Nancy J. Merrick, M.D. Address reprint requests to Dr. Brook at the Rand Corporation, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138, U.S.A.

Randomized clinical trials provide invaluable knowledge about the appropriateness of medical technologies, but they suffer from several disadvantages. They are time-consuming and expensive. As a corollary, sample sizes are limited. Thus it is impossible to draw inferences about narrowly defined groups of patients—say, patients with chronic stable angina of class 3 or 4, with two-vessel coronary artery disease, with left anterior descending involvement and very positive exercise stress test, whose angina persists despite maximal medical therapy—while at the same time including all conditions for which the procedures might be performed. In addition, randomized trials are often performed under ideal circumstances (best physicians and best hospitals) as opposed to usual circumstances, and thus their results must be generalized cautiously.

A method that avoids many of these problems is to synthesize the opinions of experts. By drawing on their knowledge and experience (some of which is derived from reading reports of randomized clinical trials), expert physicians can provide much more detailed assessments of appropriateness than even the most ambitious randomized clinical trial can produce.

We developed a method for synthesizing expert opinion and applied it to rate the appropriateness of six medical and surgical procedures under conditions that are defined with a great deal of detail. We describe that method and our application in this paper. The discussion section lists some ways in which the results might be used. For more detail on the results themselves, see references 3 and 5.

METHODS

Panel Selection

We conducted three panels. A cardiovascular panel rated indications for coronary angiography and coronary artery bypass surgery; a gastroenterological panel rated indications for cholecystectomy, upper gastrointestinal endoscopy, and colonoscopy; a cerebrovascular panel rated indications for carotid endarterectomy.

The panelists were suggested by leaders of American medicine, many of whom represented prominent medical organizations. Each panel consisted of nine physicians (various specialties—Table 1) from throughout the U.S. All panelists were

Table 1. Panel Membership

<i>Specialty</i>	<i>Panel</i>		
	<i>Cardiovascular (ANGIO, CABS)</i>	<i>Gastroenterological (CH, END, CO)</i>	<i>Cerebrovascular (CE)</i>
Medical generalists	1 family physician 2 internists	1 family physician 2 internists	1 family physician 1 internist
Medical specialists	3 cardiologists	3 gastroenterologists	2 neurologists
Surgeons	2 cardiothoracic surgeons	2 general surgeons	3 vascular surgeons 1 neurosurgeon
Radiologist	1 radiologist	1 radiologist	1 neuroradiologist

required to perform two rating tasks. The first was done before the panel meeting; the second was done at the meeting.

Initial Lists of Indications

Physicians on the project staff compiled the initial indications lists, using as guides reviews of the medical literature on each procedure. The literature reviews critically analyzed the methods used to evaluate a procedure and paid particular attention to results from randomized trials. The indications categorized patients in terms of their symptoms, past medical history, and the results of previous diagnostic tests. We attempted to compile clinical indications lists that were detailed, comprehensive, and manageable. We tried to make the indications detailed enough so that patients presenting with a particular indication would be reasonably homogeneous, in the sense that doing the procedure would be equally appropriate (or inappropriate) for all of them. We tried to make the lists comprehensive enough so that all indications for doing the procedure that arise in practice would be included. At the same time, we tried to keep them short enough so that all of the indications could be rated by the panelists within a reasonable length of time. The total number of initial indications was coronary angiography, 205; coronary artery bypass graft surgery, 370; cholecystectomy, 192; endoscopy, 1685; colonoscopy, 1086; and carotid endarterectomy, 675.

Initial Ratings

We sent the panelists literature reviews, rating sheets, and instructions. The literature reviews gave all panelists equal access to a central core of relevant literature (3). The ratings sheets listed all of the indications for each procedure and provided space for an appropriateness rating on a scale of 1 to 9. As an example, Figure 1 shows one page from the initial rating sheets for coronary angiography.

The instructions asked the panelists to rate the appropriateness of each indication as of 1981 using their own best clinical judgment (rather than their perceptions of what other experts might say), and considering an average group of patients presenting to an average U.S. physician who performed the procedure during 1981. "Appropriate" was defined to mean that the expected health benefit (i.e., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity—not necessarily in order of importance) exceeded the expected negative consequences (i.e., mortality, morbidity, anxiety of anticipating the procedure, pain produced by the procedure, time lost from work) by a sufficiently wide margin that the procedure was worth doing. "Inappropriate" meant the opposite—the negative consequences outweighed the expected benefits. Extremely appropriate indications should be rated 9, equivocal indications (neither clearly appropriate nor clearly inappropriate) should be rated 5, and extremely inappropriate indications should be rated 1. The instructions asked the clinicians to exclude cost in judging appropriateness. We did not want the panel to conclude that the procedure was not worthwhile on a poor as opposed to middle-class person simply because the person was poor. The instructions also included definitions of important clinical terms (such as a positive exercise stress test or angina) used in the indications lists. We know from conversation with the panelists that to complete the initial ratings required about 20 to 35 minutes per 100 indications.

Figure 1. Form Used for Initial Ratings of Coronary Angiography Appropriateness Scale

	1 2 3 4 5 6 7 8 9	
	1 = extremely inappropriate	
	5 = equivocal (neither clearly appropriate nor clearly inappropriate)	
	9 = extremely appropriate	
		Rating of appropriateness (circle one)
1. Asymptomatic patients		
A. Coronary angiography (CA) is indicated in patients in high risk occupations if:		
1. No exercise ECG, no exercise thallium scan, and no exercise MUGA		1 2 3 4 5 6 7 8 9
2. Negative exercise ECG and		
a. No or negative exercise thallium scan regardless of MUGA results, if any		1 2 3 4 5 6 7 8 9
b. Reversible defect on exercise thallium scan and No exercise MUGA		1 2 3 4 5 6 7 8 9
Negative exercise MUGA		1 2 3 4 5 6 7 8 9
Positive exercise MUGA		1 2 3 4 5 6 7 8 9
3. Positive exercise ECG and		
a. No exercise thallium scan regardless of MUGA results, if any		1 2 3 4 5 6 7 8 9
b. Negative exercise thallium scan and No exercise MUGA		1 2 3 4 5 6 7 8 9
Negative exercise MUGA		1 2 3 4 5 6 7 8 9

Panel Meetings

The cardiovascular panel met for two days. After brief preliminaries, they spent one day discussing and rerating coronary angiography, and one day on coronary artery bypass surgery. The gastrointestinal panel met for two and a half days. They spent one half day on cholecystectomy and two days on endoscopy, running out of time before they got to colonoscopy. Rerating of colonoscopy was done by mail without group discussion two months after the meeting. The cerebrovascular panel spent one day discussing and rerating carotid endarterectomy. The discussion of each procedure was led by the physician who prepared the initial indications list for that procedure. He or she was assisted by other physicians and social scientists.

The panelists discussed the indications for each procedure one chapter at a time. A chapter represented a clinical group of patients on whom the procedure could be performed (e.g., patients with chronic stable angina as opposed to those with chest pain of uncertain origin). During the discussion, the panelists had in front of them computer printouts that summarized their initial ratings for that chapter. For example, Figure 2 shows one page from the printout for angiography.

Figure 2. Form Used for Final Ratings of Coronary Angiography Appropriateness Scale

1 2 3 4 5 6 7 8 9

1 = extremely inappropriate
 5 = equivocal (neither clearly appropriate nor clearly inappropriate)
 9 = extremely appropriate
 ^ = number of panelists assigning this rating.
 † = this panelist's own rating.

Rating of appropriateness (circle one)

1. Asymptomatic patients

A. Coronary angiography (CA) is indicated in patients in high risk occupations if:

1. No exercise ECG, no exercise thallium scan, and no exercise MUGA 9
1 2 3 4 5 6 7 8 9
^

2. Negative exercise ECG and

a. No or negative exercise thallium scan regardless of MUGA results, if any 9
1 2 3 4 5 6 7 8 9
^†

b. Reversible defect on exercise thallium scan and

No exercise MUGA 1 1 2 2 2 1
1 2 3 4 6 7 8 9
^

Negative exercise MUGA 1 1 2 2 2 1
1 2 3 4 5 6 7 8 9
^

Positive exercise MUGA 2 1 6
1 2 3 4 5 6 7 8 9
^

3. Positive exercise ECG and

a. No exercise thallium scan regardless of MUGA results, if any 2 3 3 1
1 2 3 4 5 6 7 8 9
^

b. Negative exercise thallium scan and

No exercise MUGA 3 2 1 2
1 2 3 4 5 6 7 8 9
^

Negative exercise MUGA 3 2 2 1
1 2 3 4 5 6 7 8 9
^

By looking at the printouts, the panelists could see the distribution of their initial ratings. The numbers above the 1-to-9 rating line showed how many panelists assigned each rating. For example, all nine panelists assigned a rating of 1 to the first coronary angiography indication in Figure 2. Each panelist received a different printout; the distribution of ratings was the same on all, but the caret below the rating line showed the particular panelist's own initial rating. For example, the panelist whose printout is shown in Figure 2 rated the first two indications 1 and the third indication 6.

For all procedures except colonoscopy (which was not discussed at the panel meeting), the indications lists were substantially revised during discussions at the meetings. The changes were meant to tailor the indications so that they better described clinically homogenous patient categories. Many of the changes simply split one indication into two. For example, indications for carotid endarterectomy that described a 50 through 99 percent occlusion of the ipsilateral artery were divided into separate indications for 50 through 69 percent and 70 through 99 percent occlusion. Other changes adjusted the boundaries between indications. For example, many indications on the initial list for coronary artery bypass surgery distinguished patients who had an ejection fraction under 30 percent from those whose ejection fraction was 30 percent or greater. (An ejection fraction is the percentage of blood in the heart that is ejected in each heartbeat). For many of the final indications, the corresponding distinction was: ejection fraction under 20 percent, ejection fraction 20 to 49 percent, and ejection fraction 50 percent or greater.

In other cases, some indications were dropped and others were added. Some initial indications were divided into multiple final indications. Other groups of initial indications were merged into a single final indication. Many of the changes were even more complex transformations. In some cases, whole chapters were divided, combined, or eliminated. The final chapter titles and the number of indications in each chapter for each of the six procedures are in Table 2. The total number of final indications ranged from 196 for cholecystectomy to 2862 for colonoscopy.

The printouts showing distributions of initial ratings also served as rating sheets. After discussing each chapter, the panelists marked their final ratings directly on the printouts. For colonoscopy, however, rating sheets were mailed to the panelists eight weeks after the meeting and were returned within twelve weeks of the meeting.

Measures Used to Rate Indications

Our 1-through-9 scale is an ordinal scale. It ranks the excess or deficiency of benefit compared to risk. A 9 is always more appropriate than an 8, and an 8 is more appropriate than a 7. But risk-benefit levels are not specified for each point on the scale, so we cannot assert that the difference between a 9 and an 8 is necessarily the same as the difference between an 8 and a 7. This suggested that we should avoid measures like means and standard deviations that treat the intervals as though they were equal.

Instead, for each indication, we used the median to measure the central tendency of the nine panelists' ratings and special measures of agreement and disagreement to indicate the dispersion of the ratings.

Table 2. Final Indications for the Use of Six Medical and Surgical Procedures

<i>Chapter Number</i>	<i>Chapter Heading</i>	<i>Number of Indications</i>
Coronary Angiography		
	1. Asymptomatic patients	28
	2. Chest pain of uncertain origin	30
	2a. Chest pain of uncertain origin in males of any age or females over 50	
	2b. Chest pain of uncertain origin in females 50 or under	
	3. Chronic stable angina	108
	4. Unstable angina	28
	4a. Unstable angina age less than 65	
	4b. Unstable angina age 65 or more	
	5. During acute myocardial infarction	5
	6. Within six months of an acute myocardial infarction	68
	7. Sudden death survivors	6
	8. Following coronary artery bypass graft surgery	16
	9. Other	11
	TOTAL	300
Coronary Artery Bypass Graft Surgery		
	1. Chronic stable angina	73
	2. Unstable angina	117
	3. During acute myocardial infarction	25
	4. Following myocardial infarction	144
	4a. Following transmural myocardial infarction with or without sudden death (within 6 months)	
	4b. Following subendocardial myocardial infarction with or without sudden death (within 6 months)	
	5. Following coronary artery bypass graft surgery	91
	6. Survivors of sudden cardiac death without a myocardial infarction	4
	7. Intractable ventricular arrhythmias	11
	8. Asymptomatic coronary artery disease or chest pain of uncertain origin	14
	9. Patients undergoing primary valvular surgery	9
	TOTAL	488
Cholecystectomy		
	1. No comorbidity	49
	2. Low comorbidity	49
	3. Medium comorbidity	49
	4. High comorbidity	49
	TOTAL	196
Upper Gastrointestinal Endoscopy		
	1. Asymptomatic patients	54
	2. Follow-up to prior endoscopy	186
	3. Follow-up to a prior upper gastrointestinal series	63
	4. Upper gastrointestinal bleeding but with no evidence for portal hypertension or stigmata of liver disease	49
	5. Portal hypertension or stigmata of liver disease	12
	6. Stool positive for occult blood	288
	7. Hematochezia	152
	8. Peptic symptoms	120
	9. Obstruction, vomiting or positive saline loading test	44
	10. Dysphagia	18

(cont.)

Table 2. (cont.)

<i>Chapter Number</i>	<i>Chapter Heading</i>	<i>Number of Indications</i>
11.	Odynophagia	5
12.	Unexplained persistent chest pains after negative cardiovascular work-up	51
13.	Therapy	9
14.	Miscellaneous	18
	TOTAL	1069
Colonoscopy in Patients 16 Years and Older		
1.	Asymptomatic adults	43
2.	Asymptomatic patients undergoing cancer surveillance	339
3.	Lower bright red gastrointestinal bleeding	864
4.	Stool positive for occult blood	672
5.	Melena	324
6.	Lower gastrointestinal symptoms and signs (non-bleeding)	192
7.	Abnormal barium enema	148
8.	New diagnosis of inflammatory bowel disease	96
9.	Miscellaneous	27
10.	Follow-up to a prior colonoscopy	144
11.	Therapy	13
	TOTAL	2862
Carotid Endarterectomy		
1.	Carotid transient ischemic attack and/or amaurosis fugax—single episode	66
2.	Carotid transient ischemic attacks and/or amaurosis fugax—multiple episodes, never tried on medical therapy	66
3.	Carotid transient ischemic attacks and/or amaurosis fugax—multiple episodes with at least one recurrence since initiation of medical therapy	66
4.	Carotid transient ischemic attacks and/or amaurosis fugax—no recurrence since initiation of medical therapy (at least 3 months of therapy)	66
5.	Vertebrobasilar transient ischemic attack(s)	66
6.	Post-atherothrombotic stroke	66
7.	Stroke-in-evolution	66
8.	Crescendo transient ischemic attacks—carotid transient ischemic attacks	66
9.	Asymptomatic	90
9a.	Asymptomatic, normal stroke risk	
9b.	Asymptomatic, high stroke risk	
10.	Asymptomatic, patient to undergo surgery	180
10a1.	Asymptomatic, patient to undergo intra-abdominal or intra-thoracic surgery (except coronary artery bypass graft surgery) normal stroke risk	
10a2.	Asymptomatic, patient to undergo intra-abdominal or intra-thoracic surgery (except coronary artery bypass graft surgery) high stroke risk	
10a1.	Asymptomatic, patient to undergo coronary artery bypass graft surgery) normal stroke risk	
10a1.	Asymptomatic, patient to undergo coronary artery bypass graft surgery) high stroke risk	
11.	Dementia of vascular origin	66
	TOTAL	864

Agreement and Disagreement

In defining agreement and disagreement, we adopted the following premises. (a) We seek definitions that most people would accept as reasonable. Consequently, we try hard not to include too much in the definitions. Some people might favor broader definitions, but we hope that few would advocate stricter ones. (b) Widely acceptable definitions of agreement and disagreement will leave a middle grey area that is not obviously one or the other. (c) Our 9-point scale meaningfully divides into three 3-point regions. Ratings from 1 to 3 say that risks outweigh benefits and the procedure should not be done. Ratings from 4 to 6 say that risks and benefits are roughly equal and doing the procedure is questionable. Ratings from 7 to 9 say that benefits outweigh the risks and the procedure should be done. (d) In an imperfect world, we may need to be content with less than perfect unanimity. Although we base half of our definitions on the ratings of all nine panelists, for the other half we discard one rating at either extreme and use only the remaining seven.

We used two “conceptions” of agreement and disagreement. Both conceptions could be applied using either nine ratings or seven ratings, yielding a total of four definitions of each.

For agreement, the first, stricter conception was: The raters agreed if all of the ratings were within a single three 3-point region—1 to 3, 4 to 6, or 7 to 9. We would interpret this to mean that all of the raters agreed to one of the following statements: the procedure should not be done, doing it is questionable, or it should be done.

The second conception was somewhat more relaxed: The raters agreed if all of the ratings were within any 3-point range, even if that range straddled the boundary between two of the regions specified above. If all of the ratings were within the range 3 to 5, for example, there was agreement according to this second conception, but not according to the first. Four definitions of agreement resulted:

- A9S: All nine of the ratings fell within a single 3-point region—1 to 3, 4 to 6, 7 to 9.
- A9R: All nine of the ratings fell within any 3-point range.
- A7S: After discarding one extreme high and one extreme low rating, the remaining seven ratings all fell within a single 3-point region—1 to 3, 4 to 6, or 7 to 9.
- A7R: After discarding one extreme high and one extreme low rating, the remaining seven ratings all fell within any 3-point range.

(The codes for each definition are mnemonic: “A” for agreement, “9” or “7” for the number of ratings used, “S” or “R” for the stricter or more relaxed conception.)

As with agreement, our conceptions of disagreement could be used with either nine or seven ratings to yield four definitions. For the first conception, the raters disagreed if at least one assigned a rating of 1 and at least one assigned a rating of 9.

The second conception was more relaxed: The raters disagreed if at least one rating fell in the lowest 3-point region (1 to 3) and at least one in the highest (7

to 9). We would interpret this to mean that one thought the procedure should not be done for this indication and one thought that it should be.

Four definitions of disagreement resulted:

D9S: Considering all nine ratings, at least one was a 1 and at least one was a 9.

D9R: Considering all nine ratings, at least one fell in the lowest 3-point region (1 to 3) and at least one fell in the highest (7 to 9).

D7S: After discarding one extreme high and one extreme low rating, at least one of the remaining seven ratings was a 1 and at least one was a 9.

D7R: After discarding one extreme high and one extreme low rating, at least one of the remaining seven ratings fell in the lowest 3-point region (1 to 3) and at least one fell in the highest (7 to 9).

A Categorization of Rated Indications

How might one use the panel results to decide if a procedure was done appropriately for a particular indication? We suggest the following: each indication can fall in one of three categories—clearly appropriate, equivocal, or clearly inappropriate. We classify the indications as “equivocal” for either of two reasons: The benefits and risks of doing the procedure are roughly the same (a median rating of 4 to 6), or the panelists disagreed on the proper rating (according to one of the definitions discussed above). An indication is “clearly appropriate” if the panelists assigned a median rating in the 7-to-9 range without disagreement, and it is “clearly inappropriate” if they assigned a 1-to-3 rating without disagreement.

DISCUSSION

Our applications of the panel method demonstrated that it works. It can produce appropriateness ratings for a large number of indications for performing medical and surgical procedures in a much shorter time and at much lower cost than a randomized clinical trial.

Clearly, though, our panel method is a supplement to, not a substitute for, more traditional means of technology assessment. The panel method can synthesize existing knowledge, but it cannot create new knowledge. The results of randomized trials informed the panels’ detailed ratings. So did the panelists’ own experience and their observation of the experience of others. The panel method is a way to bring together various sources of knowledge in order to assess procedure use. The acceptance of the panel’s results by a large sample of practicing physicians has not been formally tested. The criteria have been shown, however, to a number of physicians in various parts of the United States, and the ratings have been felt to be reasonable.

We learned that it is not necessary to force perfect agreement among the panelists; there is something to be learned from disagreement as well. Use of a procedure for particular indication may be controversial for either of two reasons—the benefits and risks of doing the procedure are evenly balanced in that situation, or there is substantial disagreement among expert physicians about the balance of benefits and risks.

Panel results can be used (a) to evaluate practice patterns, (b) to aid clinical judgment, and (c) to indicate areas of uncertainty.

We will use the panel results to illuminate the clinical reasons for geographical variations in the use of three procedures (1,2,4,6,7,8). We will abstract medical records in both high-use and low-use areas of the U.S. We want to see whether procedures are used more often in some areas for indications rated inappropriate or equivocal than they are in others.

Individual physicians might use panel ratings as an aid in their own practices. If a panel rated an indication inappropriate, the physician might seriously question the advisability of doing a procedure for that indication. If the rating were equivocal, the physician might pause to ask what unusual circumstances justify use of the procedure in this particular case.

Indications which raise disagreement among expert panelists highlight areas wherein present knowledge is inconclusive. If these are indications that occur frequently in practice, they become promising candidates for further technology assessment by means of randomized clinical trials.

REFERENCES

1. Brook, R. H., Lohr, K. N., Chassin, M., et al. Geographic variations in the use of services: Do they have clinical significance? *Health Affairs*, 1984, 3, 62-73.
2. Chassin, M. R., Brook, R. H., Park, R. E., et al. Variations in the use of medical and surgical services by the Medicare population. R-3288-CWF/PMT/HF/RWJ, The Rand Corporation, Santa Monica, Cal., forthcoming.
3. Fink, A., Ed. A review of the literature and ratings of the appropriateness of indications for selected medical and surgical procedures. Series R-3204-CWF/PMT/HF/RWJ, The Rand Corporation, Santa Monica, Cal., forthcoming.
4. McPherson, K., Strong, P. M., Jones, L., & Britton, B. J. Do cholecystectomy rates correlate with geographic variations in the prevalence of gallstones? *Journal of Epidemiology and Community Health*, 1985, 39, 179-182.
5. Park, R. E., Fink, A., Brook, R. H., et al. Physician ratings of appropriate indications for six medical and surgical procedures. Series R-3280-CWF/PMT/HF/RWJ, The Rand Corporation, Santa Monica, Cal., forthcoming.
6. Shaffer, E. A. Gallstone formation and asymptomatic cholelithiasis, *Annals RCPSC*, 1985, 18, 309-315.
7. Wennberg, J. E., & Gittlesohn, A. M. Small area variations in health care delivery. *Science*, 1973, 182, 1102-1108.
8. Wennberg, J. E., Gittlesohn, A., & Shapiro, N. Health care delivery in Maine I: Patterns of use of common surgical procedures. *Maine Medical Association*, 1975, 66, 123-130.

RAND/N-3376-HHS