

This document reports the methods and results of an expert panel process convened to develop ratings of appropriateness of hysterectomy for non-emergency, non-oncology indications. The ratings were developed as part of a project to evaluate alternative methods for improving clinical decisionmaking regarding hysterectomy, and served as the basis for a set of clinical recommendations developed by a working group consisting of four obstetrician gynecologists who were members of the expert panel and a representative of the American College of Obstetricians and Gynecologists.

METHODS

The indications and ratings for hysterectomy presented in this document reflect the findings of a nine-member panel of physicians who rated the indications twice.¹ The initial ratings of appropriateness were made individually and without group discussion. The second round of ratings were made during a two-day meeting, within the framework of a modified Delphi method that is often used to bring groups closer to consensus or agreement. The group discussion focused on indications for which there was disagreement in the initial round (Park et al., 1986).

The panelists were nationally recognized physicians who were nominated by their respective specialty societies: the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians, the American College of Physicians, and the Society of General Internal Medicine. Panelists were chosen because of their clinical expertise and geographic diversity. Three of the four major census regions were represented: two panelists were from the western United States, three from the northeast, and four from the south. Three panelists were in community practice; six had academic appointments. Further, they represented different specialties: Panel members consisted of two family practitioners, two internists, and five obstetrician/gynecologists.

¹For a complete explanation of the panels and rating process, see R. E. Park et al., *Physician Ratings of Appropriate Indications for Six Medical and Surgical Procedures*, RAND, R-3280-CWF/PMT/HF/RWJ, 1986. See the Acknowledgments for a list of members of the hysterectomy panel.

Initial Indications List

Physicians on the RAND project staff compiled the initial indications list, based on a detailed literature review² and the results of a previously convened national panel. The indications categorized patients in terms of their age, symptoms, history, and the results of previous diagnostic tests. Project staff attempted to compile a detailed, comprehensive, and manageable set of indications. Indications were detailed enough so that groups of patients presenting with a particular indication would be reasonably homogenous, in the sense that performing the procedure would be equally appropriate (or inappropriate) for all members of the group. Staff attempted to make the categories comprehensive enough so that all indications for doing the procedure that might arise in practice would be included. The entire set was short enough that all of the indications could be rated by the panelists within a reasonable length of time.

The indications were organized into “chapters,” which in most cases corresponded to major symptoms or primary problems. The chapter titles and numbers from the initial indications list for hysterectomy were:

1. Cervical Dysplasia
2. Endometriosis
3. Abnormal Uterine Bleeding
4. Asymptomatic Leiomyomata
5. Leiomyomata and Bleeding (but no pain/discomfort)
6. Leiomyomata and Pain/Discomfort (but no abnormal bleeding)
7. Leiomyomata, Bleeding and Pain/Discomfort
8. *This chapter was not used*
9. Postmenopausal Patients with Leiomyomata
10. Pelvic Pain and Adhesions
11. Dysmenorrhea
12. Chronic, Non-cyclic Pelvic Pain
13. Endometrial Hyperplasia
14. Pelvic Relaxation with No Urinary Incontinence, No Pelvic Pressure/Pain
15. Pelvic Relaxation and Urinary Incontinence
16. Pelvic Relaxation and Pelvic Pressure/Pain
17. Unilateral Adnexal Mass

²S. J. Bernstein et al., *Hysterectomy: A Review of the Literature on Indications, Effectiveness, and Risks*, RAND, MR-592/2, 1997.

18. Family History of Ovarian Carcinoma

19. Miscellaneous

Chapter 8 had been dropped by a prior RAND panel and the original chapter number structure was maintained for analytic reasons.

Initial Ratings

The literature review, indications list, and instructions for rating the indications were sent to the panelists. The literature review gave all panelists equal access to a central core of relevant literature. The ratings sheets provided space for an appropriateness rating on a scale from 1 to 9. Figure 1 shows one page from the initial rating sheets.

The instructions asked the panelists to rate the appropriateness of each indication using their own best clinical judgment (rather than their perceptions of what other experts might say), and considering an average group of patients presenting in 1993 to an average U.S. physician who performed the procedure. “Appropriate” was defined to mean that the expected health benefit (i.e., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) 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. 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. Cost was to be excluded from these ratings.

Panel Meeting

The panelists met for two days in Santa Monica, California, on June 14 and 15, 1993. They discussed areas of disagreement, modified the list of indications, and once again rated the indications. The discussion was led by the physician who was primarily responsible for the literature review and assisted in the development of the list of indications. He was assisted by other physicians and social scientists on the project staff.

The hysterectomy panel discussed the indications one chapter at a time. During the discussion, each panelist had an individualized computer summary of the initial ratings for that chapter. Figure 2 shows one page from the printout. By looking at the printout, the panelists could see the distribution of initial ratings. The numbers above the 1-to-9 rating line show how many panelists assigned each rating. For example, eight panelists assigned a rating of 1 to the first indication in Figure 2, and one panelist assigned a rating of 2. Each panelist received a customized printout; the distribution of ratings was the same on all reports, but the caret (^) below the rating line showed the initial rating by a particular panelist. For example, the panelist whose report is shown in Figure 2 rated the first indication “1,” the second indication “1,” the seventeenth indication “1,” and the eighteenth indication “2.” This procedure preserved the confidentiality of individual panelists’ ratings while allowing each

panelist to see his own rating compared to the distribution of the entire group's ratings.

The hysterectomy indications list was substantially revised during the rating process. The changes were all designed to tailor the indications so that they better fit the panelists' perceptions of clinically relevant categories. The total number of indications increased by only 228, from 2,104 on the original list to 2,332 on the final list.

After discussion of each chapter, the panelists marked their final ratings directly on the printouts. The final indications for hysterectomy had the following chapter headings:

1. Cervical Dysplasia
2. Endometriosis
3. Abnormal Uterine Bleeding³
4. Asymptomatic Leiomyomata
5. Leiomyomata and Bleeding (but no pain/discomfort)
6. Leiomyomata and Pain/Discomfort (but no abnormal bleeding)
7. Leiomyomata with Bleeding and Pain/Discomfort
8. *This chapter was not used*
9. Postmenopausal Patients with Leiomyomata
10. Pelvic Pain and Adhesions
11. Dysmenorrhea
12. Chronic, Non-cyclic Pelvic Pain
13. Endometrial Hyperplasia
14. Pelvic Relaxation with No Urinary Incontinence, No Pelvic Pressure/Pain
15. Pelvic Relaxation and Urinary Incontinence
16. Pelvic Relaxation and Pelvic Pressure/Pain
17. Unilateral Adnexal Mass
18. Family History of Ovarian Carcinoma
19. Miscellaneous

The final chapter structure was identical to the original structure.

³The terminology used in this chapter reflects what the panel actually rated. Throughout the literature review, "abnormal" or "dysfunctional" uterine bleeding is referred to as "recurrent" uterine bleeding.

Chapter 1 NEUTROPHIL IS INDICATED IN PATIENTS WITH CERVICAL DYSPLASIA WHO ASKED THEM:	DEGREE OF DYSPLASIA		

	CIN I or II	CIN III/IV	
A. DO NOT WANT FUTURE PREGNANCY	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(1- 2)
B. DO NOT WANT FUTURE PREGNANCY, ARE < 40 YEARS OLD, WITH NO CHILDREN, AND WANT:			
1. No prior conization or excision	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(3- 4)
2. One prior conization or excision performed with clear margins of resection:			
a. No recurrence	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(5- 6)
b. Recurrence 2 or more years after conservative procedure	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(7- 8)
c. Recurrence < 2 years after conservative procedure	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(9- 10)
3. One prior conization or excision performed with margins of resection showing dysplasia:			
a. No repeat sampling or no dysplasia on repeat sampling	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(11- 12)
b. Repeat sampling shows dysplasia	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(13- 14)
4. Two or more prior conizations or excisions performed with clear margins of resection on last procedure:			
a. No recurrence	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(15- 16)
b. Recurrence 2 or more years after conservative procedure	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(17- 18)
c. Recurrence < 2 years after conservative procedure	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	(19- 20)

Appropriateness scale: 1 = extremely inappropriate, 5 = equivocal, 9 = extremely appropriate.

Figure 1—Sample Page from the Initial Appropriateness Rating Form

CLASSIFY 1 PREFERRENCES IS INDICATED BY PATIENTS WITH CERVICAL DYSPLASIA WHO STATE THEY:	DEGREE OF DYSPLASIA														
	CIN I or II							CIN III/CIS							
	0	1	2	3	4	5	6	7	8	9	0	1	1	1	
A. DO WANT FUTURE PREGNANCY	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 1= 21
M. DO NOT WANT FUTURE PREGNANCY, ARE < 40 YEARS OLD AND HAVE:															
1. No prior excision or excision															
- No children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 3= 43
- Children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 5= 61
2. One prior excision or excision performed with clear margins of resection															
a. No recurrence															
- No children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 7= 43
- Children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 9= 103
b. Recurrence 2 or more years after conservative procedure															
- No children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 13= 123
- Children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 13= 143
c. Recurrence < 2 years after conservative procedure															
- No children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 15= 163
- Children	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1 17= 183

Appropriateness scale: 1 = extremely inappropriate, 5 = unclear, 9 = extremely appropriate.

Figure 2—Sample Page from the Printout Used at the Panel Meeting

Analysis of the Appropriateness Ratings

The appropriateness rating scale is an ordinal one on which panelists rank their judgments about the excess of benefit over risk (including negative values when risks exceed benefits). A 9 is always more appropriate than an 8 and an 8 is always more appropriate than a 7. But risk-benefit levels are not specified for each point on the scale, so that a difference between a 9 and an 8 is not necessarily the same as the difference between an 8 and a 7. This suggests that we should avoid measures such as means and standard deviations that treat intervals as though they were equal.

The scale, however, does have some characteristics of an interval scale. The center of the scale (5) is well anchored at the point where risk equals benefit. Although the ends (1 and 9) are not precisely specified, they are anchored to some degree. At 1, risks exceed benefits by a sufficiently wide margin that the procedure should definitely not be done. At 9, it is clearly appropriate to perform this procedure.

Second, it is now established that using interval measures on ordinal scales seldom has much effect on the results.⁴ To shun interval measures entirely would throw away information. A four-point difference on our scale may not represent precisely four times as big a difference in the excess of benefit over risk as a one-point difference, but almost certainly represents a bigger difference than one point. A strictly ordinal measure would not distinguish among them.

For each indication, the median was used to measure the central tendency for the nine panelists' ratings and the mean absolute deviation from the median to measure the dispersion of the ratings. These measures are well suited, we believe, to the appropriateness scale. Table 1 describes for all ratings the panel's median and mean absolute deviation from the median, and the percentage of agreement and disagreement in the panel's initial and final appropriateness ratings. The table shows that both the ratings and the dispersion (the mean absolute deviation from the median) decreased from the initial to the final ratings.

To determine agreement and disagreement among panelists, we adopted a somewhat statistical definition so that future studies using panel compositions of other

Table 1
Mean Ratings, Dispersion of Ratings, and Extent of Agreement
and Disagreement in Panel's Initial and Final
Appropriateness Ratings for Hysterectomy
(rated by indication category)

Item	Initial Ratings	Final Ratings
Number of indications	2,105	2,332
Panel median	2.60	2.78
Mean absolute deviation from median	1.38	0.90
Percentage agreement	48.1	60.5
Percentage disagreement	9.8	2.7

⁴See, for example, L. Moses, J. D. Emerson, and H. Hosseini, "Analyzing Data from Order to Categories," *New England Journal of Medicine*, 1984, Vol. 310, pp. 442-448.

than nine members could easily adapt our definitions of agreement and disagreement; the definition also makes the treatment of missing ratings easier. In this approach, we frame the definitions as tests of hypotheses about the distribution of ratings in a hypothetical population of repeated ratings by similarly selected panelists.

For agreement, we test the hypothesis that 80 percent of the hypothetical population of repeated ratings are within the same region (1–3, 4–6, 7–9) as the observed median rating. If we are unable to reject the hypothesis on a binomial test at the 0.33 level, we say that the indication is rated “with agreement.” For nine ratings, this definition of agreement requires that no more than two of the ratings be outside the three-point region that contains the median.

For disagreement, we test the hypothesis that 90 percent of the hypothetical population of repeated ratings are within one or two extra-wide regions (1–6 or 4–9). If we have to reject the hypothesis on a binomial test at the 0.10 level, we conclude that the indication is rated “with disagreement.” For nine ratings, this definition of disagreement is satisfied when three or more ratings are in the 1–3 region and three or more are in the 7–9 region. Thus, for nine ratings, the new definition is equivalent to the one used in previous publications.⁵

Using the above definitions, panelist agreement increased significantly from the initial (48.1 percent agreement) to the final (60.5 percent agreement) ratings. Disagreement also decreased significantly from 9.8 to 2.7 percent.

The most disagreement occurred in Chapter 16 (“Pelvic Relaxation and Pressure/Pain”); the panelists disagreed on 13.8 percent of the ratings. No disagreement occurred in Chapters 3, 4, 6, 12–14, and 17–19.

We recommend caution when reviewing the indications. The overall median value is not a meaningful figure. The median does not tell us anything about the extent to which hysterectomy is used appropriately in clinical practice; to determine that, actual data on how the procedures are used must be compared with the indications listed. In addition, even though we had more than 2,000 indications, there still may be meaningful differences among patients placed in any one category. This means that extenuating clinical circumstances not covered by the indications may make their application inaccurate in some patients. On a positive note, however, these indications, and their ratings, reflect a medical tradeoff of risk and benefit in which cost considerations have been removed.

RESULTS

Figure 3, preceding the final list of rated indications for hysterectomy, provides a guide to reading the ratings. The chapter headings and specific indications are shown along the top and the left margin. Each rating contains a distribution of panelist ratings along with summary statistics for that particular indication. This

⁵M. Chassin et al., “Does Inappropriate Use Explain Geographic Variations in the Use of Health Services? A Study of Three Procedures,” *Journal of the American Medical Association*, 1987, Vol. 258, pp. 2533–2537.

Appropriateness Scale	
	1 2 3 4 5 6 7 8 9
1	= extremely inappropriate
5	= uncertain (neither clearly appropriate nor clearly inappropriate)
9	= extremely appropriate

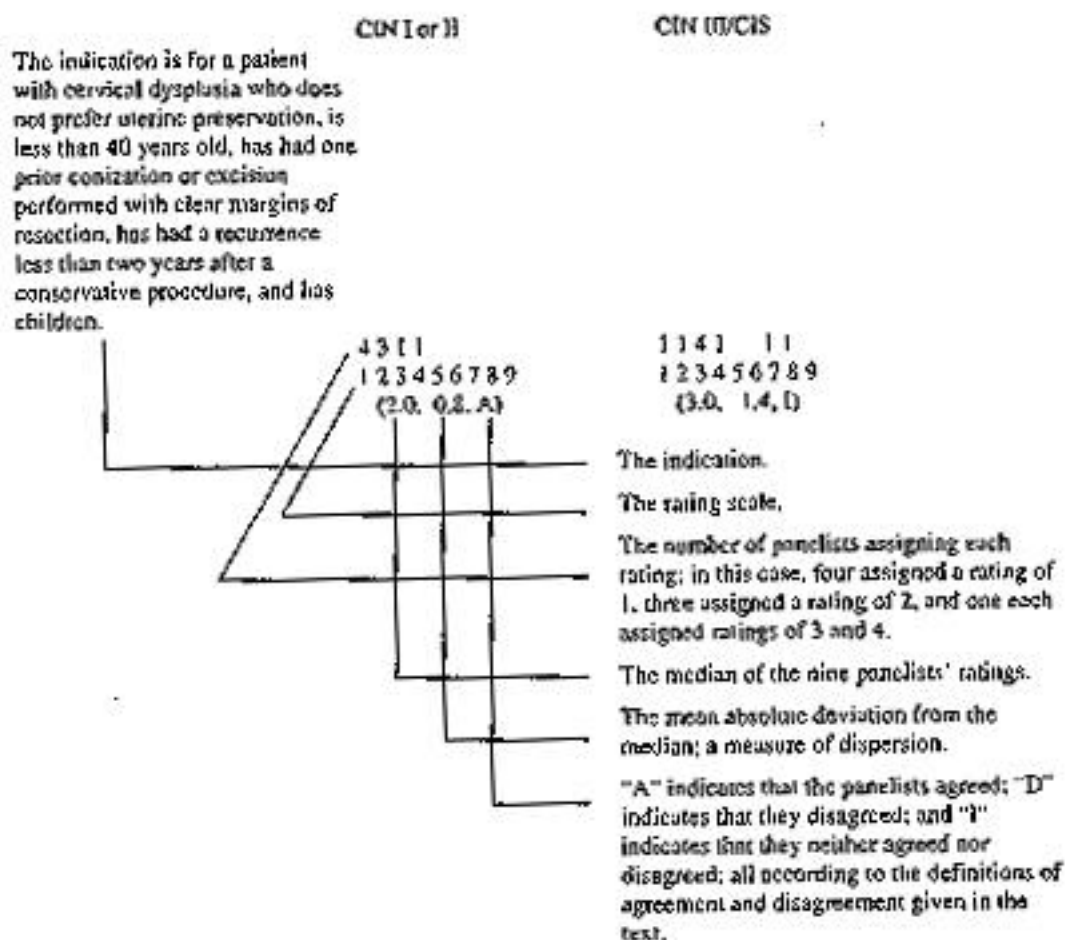


Figure 3—A Key to Reading the Final Results of Appropriateness Ratings for Each Indication

example is taken from the last pair of indications on the first page of ratings. This indication is for a patient with cervical dysplasia, who does not prefer uterine preservation, is less than 40 years old, has had one prior conization or excision performed with clear margins of resection, has had a recurrence less than two years after a conservative procedure, and has children.⁶ Cervical dysplasia is classified as either mild-to-moderate (CIN I or CIN II) or severe (CIN III or CIS) dysplasia.

The rating bar goes from 1 to 9, indicating the possible responses. Panelist ratings for this indication were dispersed: Four panelists rated this indication 1, three panelists rated this indication 2, and the remaining two panelists assigned ratings of 3 and 4. In this example, the median rating was 2 (inappropriate) and the mean absolute deviation from the median (a measure of dispersion) was 0.8. The letter “A” in this position of the table entry indicates agreement, the “I” indicates indeterminate (neither agreement nor disagreement), and a “D” indicates disagreement.

⁶Definitions of clinical terms used in the indications structure are at the beginning of each chapter in the ratings section.