Appropriateness of Upper Gastrointestinal Endoscopy: Comparison of American and Swiss Criteria

JOHN-PAUL VADER,* BERNARD BURNAND,* FLORIAN FROEHLICH,† KARINE DUPRIEZ,* TANIA LAREQUI-LAUBER,* ISABELLE PACHE,† ROBERT W. DUBOIS,‡ JEAN-JACQUES GONVERS† and ROBERT H. BROOK§

*Institute of Social and Preventive Medicine, University of Lausanne, CH-1005 Lausanne Switzerland; †Division of Gastroenterology, Policlinique médicale universitaire, University of Lausanne, CH-1005 Lausanne Switzerland; ‡Value Health Sciences, Santa Monica, CA 90404, USA; §University of California (UCLA), School of Medicine and School of Public Health, Los Angeles, CA 90024, and, RAND Corporation, Santa Monica, CA 90407, USA.

Objective: Examine the reproducibility of the RAND method for developing criteria for the appropriateness of medical procedures.

Design: Comparison of two sets of explicit criteria for appropriateness of upper gastrointestinal (UGI) endoscopy, developed by separate expert panels from two countries.

Setting: United States, Switzerland.

Study participants: National experts from different medical specialties involved in the referral or application of UGI endoscopy.

Intervention: Each panel was presented with about 500 clinical scenarios (indications) that were rated on a nine-point scale as to the appropriateness of performing UGI endoscopy for a patient with that clinical presentation.

Main outcome measures: (1) distribution of appropriateness ratings and intrapanel agreement categories between the two panels, (2) between-panel agreement of assigning appropriateness for comparable indications and, (3) percentage of indications with major between-panel differences.

Results: Ratings for 2/3 of indications could be compared. The Swiss panel showed higher intrapanel agreement (54-6% versus 46-2%, \( P = 0.002 \)). Seventy-eight per cent of comparable indications were assigned to identical categories of appropriateness by both panels (kappa = 0.76, \( P < 0.001 \)). For 93% of the 376 comparable indications, there were no major interpanel differences.

Conclusion: Separate expert panels in different countries, using a standardized methodology, produce criteria for appropriateness of medical procedures that are similar. Given the resources being invested throughout the world in developing criteria and guidelines, international collaboration in seeking optimal use of limited health care resources should be intensified. © 1997 Elsevier Science Ltd. All rights reserved.

Key words: Quality of health care, appropriateness of care, Switzerland, process assessment (health care), quality assurance (health care), utilization review, methods, reproducibility of results, gastroscopy, endoscopy (gastrointestinal).

INTRODUCTION

Both quality and cost of medical care are critically dependent on the appropriate use of medical procedures. For most clinical situations encountered, however, there are no prospective controlled clinical trials demonstrating the efficacy of upper gastrointestinal endoscopy on patient outcome; there is little hope that this situation will change in the near future. Determining appropriateness of indications for procedures must, therefore, depend on other methods. Among alternative methods that have been put forth, the RAND appropriateness method [1], which combines a detailed review of the literature with a modified Delphi approach to collective expert opinion, has been proposed and is being increasingly implemented.

When applying in Switzerland, on an experimental basis, criteria that had been developed in the United States by this method, we found that clinicians were uneasy about the imported criteria, suggesting the necessity of developing national criteria corresponding to accepted state-of-the-art practice in Switzerland [2]. For this reason, we were interested in testing the hypothesis that criteria developed by panels in different countries, but of similar composition, produce appropriateness ratings that are similar. In fact, previous international comparisons [3,4] have focused more on the results of applying, retrospectively, criteria from different countries to patients. The present study scrutinizes the criteria themselves. Also, whereas previous
comparisons [3,5] [4] have been limited to global results between panels (i.e. the mean of median ratings or distribution of appropriateness categories), this study goes beyond that by examining agreements for each of several hundred comparable clinical scenarios. Building thus on previous work in the field, the present study compares detailed and explicit criteria for appropriateness of upper gastrointestinal endoscopy, developed by national expert panels from two countries.

METHODS

Two multi-specialty expert panels were convened in 1994, one in March in the United States (US) and one in October in Switzerland to determine the appropriateness of indications for upper gastrointestinal (UGI) endoscopy. Both panels had a similar composition in terms of specialty representation: five gastroenterologists, two internists, one general practitioner and one surgeon. This composition was selected to reflect the typical use of, and decision-making for, the procedures in the two countries. Both panels followed the standardized procedure [1,6] for the RAND Appropriateness Method, which is hereafter briefly described.

Based on an extensive literature review, a rating matrix of all potential indications for upper GI endoscopy was designed. Table 1 shows the major categories of indications used in the rating process.

Other elements used to render indications clinically specific included: patient age, symptoms, first or recurrent episode, duration of and response to previous medical treatment, Helicobacter pylori status and treatment, previously documented GI pathology (type and grade of pathology, type of examination, date since last examination) and use of non-steroidal anti-inflammatory drugs. Combinations of all relevant items were designed to cover virtually all imaginable indications for which upper GI endoscopy might be considered.

Each clinically specific situation was first rated on a 9-point scale (1 = extremely inappropriate, 9 = extremely appropriate, 5 = equivocal/uncertain) separately by the nine panelists. Following that, panelists came together and were provided with reports showing their initial ratings and the anonymous distribution of other panelists' ratings. Indications were discussed and panelists re-assigned ratings of appropriateness to each indication. It was this second (final) rating which was used to compare the two panels.

To facilitate comparisons, the 9-point scale was consolidated into appropriate, equivocal and inappropriate, using the median value of the panel rating (1-3 = inappropriate; 4-6 = equivocal; 7-9 = appropriate), and the degree of agreement among the panelists (i.e. all indications where there was disagreement were classified as Equivocal, irrespective of the median score).

Intrapanel agreement was present when, after discarding one extreme high and one extreme low rating, the remaining seven ratings all fell within any single 3-point region. Disagreement was considered to be present when, after discarding one extreme high and one extreme low rating, at least one of the seven remaining ratings fell in the lowest 3-point region (1-3) and at least one in the highest (7-9).

The primary objective of the panel was to develop criteria rather than to test the reliability of the process. Therefore, neither panel was bound by the initial proposed indication structure of indications and was free to modify it during the panel process to ensure clinical specificity of the scenarios that were discussed. For this reason, not all indications are identical between the two panels. The Swiss panel began with a modified version of the indication matrix that had been developed by the US panel, adapting and modifying it to correspond to clinical reasoning in Switzerland. Main analyses were based on comparable indications that were either exactly identical in wording or sufficiently similar to make comparisons reasonable. Similarity of indications between the two panels was judged independently by two of the authors (FF and TLL) who arrived at the same conclusion as to which of the indications were similar. Table 2 shows extracts from formulations that were judged similar and the number of indications covered by those formulations. It also presents an example of a full indication, showing the respective US and Swiss formulations that were judged similar.

Outcome measures include the number of comparable indications, the distribution of appropriateness ratings and intrapanel agreement categories between the two panels, for all indications and for those that were comparable, the between-panel agreement for comparable indications and the percentage of comparable indications for which major differences were encountered. The latter is here defined as at least a 3-point
### RESULTS

The overall number of indications rated was similar (US: 549; Swiss: 598). For all indications, there was no statistically significant difference in the distribution of rating categories (1–3, 4–6, 7–9) \( (P = 0.6) \), or degree of intrapanel agreement \( (P = 0.7) \).

For 82 indications, the wording of the indications was exactly identical in both the Swiss and the US panels; 294 indications had minor differences in content (were similar). Three hundred and seventy-six indications exactly identical in both the Swiss and the US panels; 294 had minor differences in content (were similar). For 82 indications, the wording of the indications was exactly identical in both the Swiss and the US panels; 294 indications had minor differences in content (were similar).

There was no difference in the distribution of rating categories \( (P = 0.6) \). On the other hand, a higher percentage of indications showed agreement within the Swiss panel \( (54.6\% \text{ vs } 46.2\%) \) with a corresponding lower percentage of disagreement \( (8.1\% \text{ vs } 16.0\%) \); this difference was statistically significant \( (P = 0.002) \).

Table 3 shows agreement using the trichotomized appropriateness categories (appropriate, equivocal, inappropriate) between the two panels for comparable indications. When panels rated comparable indications, 293 \( (78\%) \) were assigned to identical categories of appropriateness by both panels \( (kappa = 0.76; P < 0.001) \).

There were two indications rated appropriate by one panel and inappropriate by the other. They represent fairly infrequent clinical situations. The first is an asymptomatic patient referred for UGI endoscopy as a screening test for digestive malignancy. This hypothetical patient had had a surgical resection of a head and neck carcinoma, but has no current GI signs or symptoms. The last post-surgical endoscopy goes back more than 12 months. The Swiss panel felt that for such a patient an UGI would be inappropriate, whereas the US panel felt it would be appropriate. The second situation represents a patient being considered for a diagnostic UGI endoscopy to evaluate an esophageal or gastric ulcer seen on a prior radiographic study (upper GI series) and not evaluated by endoscopy. Here the Swiss panel felt the indication for UGI endoscopy was appropriate, whereas the US panel felt it was inappropriate. It is noteworthy, however, that upper GI series are currently used only rarely in Switzerland and the very existence of such a scenario was considered highly unlikely by the Swiss panelists. Other than these two scenarios, there were no frankly discordant appropriateness ratings between the two panels for scenarios that could be compared.
Figure 1. Distribution of median ratings of appropriateness of indications for upper gastrointestinal endoscopy and degree of intrapanel agreement. Results of comparable indications (n = 376) for US and Swiss panels.

Figure 2 shows the real difference in median ratings between the two panels for comparable ratings. For 350 (93%) of the 376 comparable indications, no major differences (> 3 points in median ratings) were encountered.

For comparable indications, the mean of the median ratings was 4.58 for the US and 4.45 for the Swiss panels (P = 0.06). For the gastroenterologists on the panels who performed the procedures (doers), the mean intrapanel (US-Swiss) difference for comparable indications was 0.31 (95% CI 0.17-0.46, P < 0.001); it was 0.11 for those who did not perform the procedures (referrers) (95% CI 0.06-0.27, P = 0.2). In both panels, doers tended to rate slightly higher, i.e. more appropriate, than referrers, though this difference was more marked in the US panel (+0.64 average median rating for the US panel for doers, compared to referrers; +0.44 for the Swiss panel).

Of 1540 patients with digestive symptoms who pre-
TABLE 4. Interpanel agreement on appropriateness of indications for upper digestive endoscopy among outpatients presenting with digestive symptoms, n = 1362

<table>
<thead>
<tr>
<th>US panel</th>
<th>Swiss panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>128</td>
</tr>
<tr>
<td>Equivocal</td>
<td>-</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>50</td>
</tr>
<tr>
<td>218</td>
<td>111</td>
</tr>
</tbody>
</table>

kappa = 0.88 (P < 0.001).

Presented to a Swiss University medical outpatient clinic, 1362 (88%) had indications for which ratings could be compared between the two panels. Table 4 shows the appropriateness ratings for those patients. For 1256 (92%) patients, the ratings of appropriateness were identical as judged by the two panels (kappa = 0.88, P < 0.001). Neither of the two discordant scenarios (i.e. judged appropriate by one panel and inappropriate by the other) identified in Table 3 were encountered among the 1540 consecutive patients presenting in this outpatient setting.

DISCUSSION

Although the RAND method is gaining acceptance and use as a means to develop criteria of appropriateness of medical procedures, little has been published about the reproducibility of criteria stemming from this process. The study presented here provides strong evidence that separate expert panels in different countries, of similar composition, using the RAND appropriateness method and given similar baseline materials (literature review and clinical scenario matrix), produce criteria for appropriateness of medical procedures that are similar.

What are the limitations of this study and what are its possible implications for optimal use of health care resources and for improving the quality of care?

The results presented here, with their high level of agreement, refer to the use of a standardized method of criteria development. It would be unreasonable to assume that high agreement would also exist with the development of criteria using methods that are less formal and rigorous and often more implicit and subjective, which is the case for most guidelines that are being developed. On the other hand, one might argue that precisely because other methods are less rigorous and less clinically specific, agreement might occur simply because the criteria are vague, general and non-specific. The fact that high agreement was present, in spite of the high specificity and detail of the clinical scenarios, lends further support to the validity of the RAND method.

There are, however, possible explanations for this high agreement, which might be considered sources of potential bias in this study. Both panels began with the same literature review which clearly described the current state of published results on the use, effectiveness and complications of endoscopy in different situations. This was provided to the members of both panels and reference to it was made throughout the discussion process. Panelists were not, however, bound by any data from the literature and were free to ignore the literature or to make reference in the discussion to literature that had not been included in the literature review that had been provided. Both panels began with a similar matrix of clinical scenarios. In fact, the Swiss panel began with the matrix that had been produced by the US panel. Panelists were, however, free to modify these matrices and did not hesitate to do so when they felt it necessary in order to evaluate indications that included all the necessary information to judge the appropriateness of the indication. It was precisely because the panels did not feel bound by the common matrix that not all indications could be compared. Finally, one of the two moderators (RWD) was common to both panels and this may have created an artificial similarity of the questions raised during the discussion. However, considering both the expertise of the panelists, their knowledge and experience and the confidential nature of the rating process, it is unlikely that these biases are able to explain the high degree of between-panel agreement.

The two panels we studied considered appropriateness in the framework of two countries with fairly similar, liberal-based health care systems, where outpatient care is provided largely by practitioners in private practice, inpatient care is provided by a public/private mix of hospitals and a free market system plays a preponderant role. This fact might limit the generalizability of our results. A previous study, comparing US and British ratings of appropriateness for cardiovascular procedures (angiography and bypass surgery) [3], showed that the UK panel rated, on average, indications as less appropriate and this shift was attributed, at least partially, to the more restrictive and regulated access to care in the UK. The values underlying a health care system will no doubt influence the rating of appropriateness, but the degree to which that is the case needs to be studied and quantified. However, two other previous studies comparing use of criteria from the US with those of Canada [4] and Israel [5] also found considerable agreement in applying those criteria to patients, in spite of the quite different health care systems involved.

Although our results point to a high degree of agreement between the two panels (78% of appropriateness ratings were identical and only 2 of the 376 that were compared were frankly discordant), the difference in the mean of the median ratings between the two panels almost reached statistical significance and the difference between the doers of the procedure (gastroenterologists) was statistically significant between the two panels. However, other than pointing to a slightly more aggressive approach to the use of UGI endoscopy among
gastroenterologists in the US compared to those in Switzerland, this statistically significant difference does not change our overall conclusions and, in fact, has little practical implication or clinical significance. This result stems from testing a difference in mean value of a fairly large number of comparisons, using a paired t-test, a powerful tool to detect statistical significance. However, the (statistically significant) average difference between the doers of the two panels was 0.3, on a scale of 1 to 9. Condensing that scale into appropriate (7–9), equivocal (4–6) and inappropriate (1–3) further reduces the practical impact of this statistical difference on the overall results of comparability of appropriateness ratings between the two panels. For this reason, we feel the categorical data in Table 3 represent a more robust and a more practical approach to comparing panel results.

Because the goal of the panel process was to develop usable criteria and not to test the comparability of criteria developed by the two panels, we were concerned that the definitions of indications (clinical scenarios) would differ so widely that comparison would not be possible. The fact that one-third of the indications could not be compared may be seen as a weakness of our results. However, the high level of agreement among those indications which could be compared, especially the fact that complete agreement is seen for such a high percentage (92%) of indications concerning real patients presenting in actual practice, point in the opposite direction: agreement appears to be strong on indications that correspond to real and frequent cases.

Although this study provides strong evidence that expert panels from different countries can produce criteria for appropriateness of medical procedures that are similar, it does not provide evidence that this will always be the case. The results of this study do not tell us if the high level of agreement will be found for panels rating other procedures. For this reason, similar studies, with criteria of appropriateness for other medical procedures are required, before our conclusions about the reproducibility of the method can be generalized.

In the meantime, however, it is not unreasonable to assume that criteria developed in one country can be of considerable benefit to other countries in evaluating and promoting the appropriate use of medical care. A recent literature review and scenario matrices prepared in one country appear to be a good starting point for the process of defining appropriateness of care in another country.

Given the time, effort and money being invested in developing criteria and guidelines and given the economic constraints bearing on all health care systems, this could lead to significant synergy and savings. For this reason, international collaboration in developing guidelines and appropriateness criteria should be intensified.

Acknowledgements: This study was supported by the Swiss National Science Foundation, grant number 32-40522.94. Robert W. Dubois, M.D., Ph.D. is an employee of Value Health Sciences (a for-profit company) which is a subsidiary of Value Health, Inc. Value Health Sciences develops practice guidelines using the methodologies described in this article. He is also a shareholder in Value Health.

REFERENCES


