International Journal for Quality in Health Care 2006; **Volume 18, Number 3:** pp. 177–182
Advance Access Publication: 7 April 2006
10.1093/intqhc/mzl005

Appropriateness of colorectal cancer screening: appraisal of evidence by experts

ISABELLE PEYTREMANN BRIDEVAUX[1,2], ANNE-MELODY SILAGHI[1], JOHN-PAUL VADER[1], FLORIAN FROEHLICH[3,4], JEAN-JACQUES GONVERS[3] AND BERNARD BURNAND[1]

[1]Health Care Evaluation Unit, [2]Health Services Research Unit, Institute of Social and Preventive Medicine, Hospices-CHUV, [3]Division of Gastroenterology & Medical Outpatient Department, Hospices-CHUV, University of Lausanne, Switzerland, and [4]Gastroenterology Department, University of Basle, Switzerland

Abstract

Objectives. To evaluate how the level of evidence perceived by an international panel of experts was concordant with the level of evidence found in the literature, to compare experts perceived level of evidence to their appropriateness scores, and to compare appropriateness criteria for colonoscopy between experts and an evidence-based approach.

Design. Comparison of expert panel opinions and systematic literature review regarding the level of evidence and appropriateness of colonoscopy indications.

Participants. European Panel on the Appropriateness of Gastrointestinal Endoscopy multidisciplinary experts from 14 European countries.

Main outcome measures. Concordance and weighted kappa coefficient between level of evidence as perceived by the experts' and that found in the literature, and between panel- and literature-based appropriateness categories.

Results. Experts overestimated the level of published evidence of 57 indications. Concordance between the level of evidence perceived by the experts and the actual level of evidence found in the literature was 36% (weighted kappa 0.18). Indications for colonoscopy were reported to be appropriate, uncertain, and inappropriate by the experts in 54, 19, and 27% of the cases, and by the literature in 37, 46, and 17% of the cases. A 46% agreement (weighted kappa 0.29) was found between literature-based and experts' appropriateness criteria.

Conclusions. Experts often overestimated the level of evidence on which they based their decisions. However, rarely did the experts' judgement completely disagree with the literature, although concordance between panel- and literature-based appropriateness was only fair. A more explicit discussion of existing evidence should be undertaken with the experts before they evaluate appropriateness criteria.

Keywords: colonoscopy, colorectal cancer, concordance, level of evidence, RAND appropriateness method, weighted kappa

Clinical practice guidelines (guidelines) may contribute to enhance quality of care [1]. A prerequisite is that guidelines should be systematically developed and based on the best available evidence [2]. In the absence of high-quality evidence from published literature, expert opinion is often used [3]. However, expert opinion is considered to be the literature's lowest level of evidence [4], and the development process of recommendations based on expert opinion is not always explicit, leading thus to scepticism concerning their merit and usefulness. The non-systematic development of guidelines, including lack of transparency about how experts' judgement is elicited or formulated, has been acknowledged to be responsible for the variations in guidelines quality and content [5–9].

The RAND Appropriateness Method (RAM), developed during the 1980s, combines expert opinion with a comprehensive review of available scientific evidence and is the most widely accepted method of developing appropriateness criteria from which recommendations can be derived [10]. One way to validate RAM results is to compare experts' judgements to literature-based evidence. In a previous study, we observed that the degree of agreement between recommendations for the use of colonoscopy developed by Swiss experts and published evidence was moderate to good [11].

In this study, we focused our analysis on the relationship between the level of evidence found in the literature and the level of evidence as perceived by the panel of experts. The objectives of this study were, firstly, to evaluate to what extent the level of evidence perceived by an international panel of experts was in agreement with the level of evidence found in the literature, secondly, to compare the experts' perceived level of evidence to their appropriateness scores, and thirdly, to compare appropriateness criteria for colonoscopy

Address reprint requests to Isabelle Peytremann Bridevaux. Institute of Social and Preventive Medicine, 17 Bugnon, CH-1005 Lausanne, Switzerland. E-mail: isabelle.peytremann-bridevaux@hospvd.ch

between experts and an evidence-based approach. We hypothesized that the concordance between the perceived level of evidence and the level of evidence actually existing in the literature would be relatively good, because the RAM is based on an extensive literature review which should constitute a reference ground for experts' judgement and panel discussion. Since appropriateness criteria for performing colonoscopy were based on existing evidence, we also hypothesized that the concordance between panel- and literature-based criteria would be at least moderate to good (i.e. corresponding to a kappa value of 0.5–0.8).

Methods

This study used the RAM, a formal group process modified from the Delphi method [10]. Briefly, in the framework of the European Panel on the Appropriateness of Gastrointestinal Endoscopy (EPAGE), a list of all the possible clinical indications for performing both upper and lower gastrointestinal endoscopy was first developed, based on an extensive literature review and completed by information obtained from experts. This list totalled about 600 indications. Then, a panel of experts studied the literature review, considered the literature in light of their own clinical experience, discussed, and rated the appropriateness of a series of indications to perform a gastrointestinal endoscopy. Further details about the methodology of the EPAGE project can be found elsewhere [12].

In this study, only the subset of 95 colorectal cancer screening indications was considered. They related to screening for early diagnosis of colorectal cancer in asymptomatic patients ($n = 37$), in patients with inflammatory bowel disease ($n = 29$), and after polypectomy or curative-intent resection of a colorectal cancer ($n = 29$). In November 1998, a multidisciplinary panel of 14 experts (8 gastroenterologists, 2 surgeons, and 4 primary care physicians) from nine European countries met in Lausanne, Switzerland, to develop appropriateness criteria for performing colonoscopy. The ratings were confidential and took place in two rounds. The results of the first rating, done independently at home and forwarded to the organizers, were distributed to all experts during the panel meeting allowing them to compare their own ratings to those of their fellows. The second rating took place during the panel meeting, and its results were used to determine the appropriateness of each indication.

The definition of appropriateness which was used by the experts was the following: the indication to perform a medical procedure is appropriate if the expected health benefit exceeds the possible negative consequences by a sufficiently wide margin that the medical procedure is worth performing. Financial costs have no direct bearing on the appropriateness of the procedure evaluated by this method [10].

Ratings were scored on a 9-point scale (1, extremely inappropriate; 5, uncertain; and 9, extremely appropriate), and the median panel rating was used to summarize the panel's ratings. For each indication, colonoscopy was rated appropriate if the median rating was 7–9 without disagreement, and inappropriate if the median rating was 1–3 without disagreement. Otherwise, the appropriateness was considered uncertain. Disagreement was defined, as occurring in this 14-member panel, as four panellists rating in the 1–3 range and four panellists rating in the 7–9 range for the same indication [10].

A total of 226 relevant articles were retrieved from various computerized databases (Medline, Embase and Cochrane Library), analysed, and summarized in a review which was given to the experts as a framework of evidence upon which to base their judgements [12]. The literature review was reported in a series of four articles [13–16].

The literature level of evidence was evaluated based on the study design as proposed by the evidence-based medicine approach. We used a five-level (I–V) summary table derived from a published list [17]. The level of existing evidence according to the expert was measured on a four-category scale (A–D), which was developed by the experts participating in the European Commission BIOMED Concerted Action on the appropriateness of medical and surgical procedures which supported this project. When comparing the levels of evidence, the categories II, III, and IV of the literature's hierarchy of evidence were combined under the label of 'controlled and observational trials', in order to have four categories for both measures of level of evidence (Table 1). The classification of the level of evidence from the literature was done independently by two persons, and disagreements were resolved by joint review of the articles.

On the basis of the literature only, appropriateness categories (appropriate, uncertain, and inappropriate) were also defined for each of the 95 indications. An indication for colonoscopy was considered appropriate if all study results were concordant and in favour of colonoscopy, inappropriate if all study results were concordant but not in favour of colonoscopy, and uncertain otherwise. Colonoscopy indications for which we did not find evidence from the literature were considered as uncertain. We did not attribute different weights to the studies according to their design or quality. Sensitivity analysis was performed using other classification schemes of literature-based categories. Firstly, the indication was considered appropriate if 2/3 (and not all) of the studies gave concordant results and secondly, the recommendation from the article showing the highest level of evidence was considered to be the reference for the appropriateness measure.

Concordance and weighted kappa coefficients between the level of evidence as perceived by the experts' and the level of evidence found in the literature as well as between the panel- and literature-based appropriateness categories were calculated. Proportion of concordance corresponds to the ratio of the number of indications for which both the results of the panellists and of the literature were identical, divided by the total number of indications (95). Kappa is a measure of reliability for categorical measures. It is designed to correct for chance agreement and calculated as follows:

$$\mathrm{K} = \frac{P_\mathrm{o} - P_\mathrm{e}}{1 - P_\mathrm{e}}$$

where P_o is the observed concordance and P_e the concordance expected by chance alone. Weighted kappa is a variant of

Table I Scales used for the level of evidence from literature and experts' perceived level of evidence

Literature hierarchy of evidence		Experts' perceived level of evidence		Examples of study designs
I	Randomized controlled trial or systematic review	A	Reliable scientific evidence or guideline based on that kind of evidence	Randomized controlled trial
II	Clinical controlled trial (without randomization)	B	Weaker scientific evidence	Clinical controlled trial (without randomization)
III	Cohort study	B	Weaker scientific evidence	Observational studies
IV	Retrospective study	B	Weaker scientific evidence	Observational studies
V	Case report, case series, experts' opinion	C	Experts' opinion	Descriptive studies, reviews, experts' opinion
Nihil		D	Own or peers' opinion	Own opinion, no evidence from literature

kappa that gives 'full credit' when two observations agree exactly, and 'partial credit' when they disagree, depending on how far apart these two measurement are [18]. Several weighting schemes exist; we used $1 - |i-j|/(k-1)$, where i and j index the rows and columns of the ratings by the two raters, and k is the maximum number of possible ratings.

All statistical analyses were done using STATA 7.0.

Results

Concordance between the two persons who classified the level of evidence from the literature was 96% (weighted kappa 0.96) when considering the cases of early detection of colorectal cancer in patients with inflammatory bowel diseases, 49% (weighted kappa 0.41) when considering cases of early detection of colorectal cancer after polypectomy or curative-intent resection, and 47% (weighted kappa 0.46) when considering cases of early detection of colorectal cancer in asymptomatic patients.

On average, the time experts spent studying the literature review was 7.9 hours [19]. The concordance between the level of evidence as perceived by the experts and the actual level of evidence in the literature was 36% ($n = 34$), corresponding to a weighted kappa of 0.18 (Table 2). Experts overestimated the level of evidence of 57 indications (60%) and underestimated it in four indications (4%). Among the studies identified and included in the literature review, 'own opinion/experience' was never cited collectively by the experts, even though no proof existed for 14 indications. In the presence of reliable level of evidence, the proportion of indications judged to be uncertain was lowest (2/26, 8%) (Table 3).

Table 4 summarizes the concordance table, comparing the level of appropriateness according to the literature and to the experts for each of the 95 colorectal cancer screening indications. Compared with the appropriateness criteria derived from the literature, the panel of experts considered more colonoscopies to be either appropriate (54 versus 37%) or inappropriate (27 versus 17%). Uncertainty was noted in 46% of the indications according to the literature but only in 19% according to the experts. Thus, full concordance between the literature's and experts' appropriateness criteria was found in 46% of the cases ($n = 44$, weighted kappa 0.29), and full discordance occurred in only seven cases (7.4%). Sensitivity analysis showed that slightly higher weighted kappas were obtained when broader criteria of literature appropriateness were used. A weighted kappa of 0.41 was found when 2/3 of the articles agreed in their recommendations, and a weighted kappa of 0.31 when the article with the highest level of evidence was considered to be the reference.

Discussion

We examined how much the level of evidence from literature and experts' perceived level of evidence were associated. We also compared the agreement of appropriateness criteria for colonoscopy between a panel of experts and a literature-based approach. The finding that experts often overestimated the level of evidence on which they based their medical decisions should be emphasized. Even though we rarely noticed complete discordance between experts and literature, the weighted kappa coefficient indicated only fair agreement.

Why do physicians, even if considered clinical experts by their peers, overestimate the level of evidence of published literature on which they base their medical decisions? Actually, most experts were not experienced methodologists in critical literature appraisal, and grading the quality of evidence is a difficult process, which is reflected by the fact that evidence from four randomized clinical trials was not acknowledged as reliable evidence by the experts. Despite the existence of several systems for grading the level of evidence, none seemed to incorporate all the important concepts and dimensions required [4], which prompted a group of experts to develop the GRADE system [20]. However, despite having addressed key shortcomings of the above-mentioned tools, the evaluation of the GRADE approach showed many areas of disagreements, underscoring the complexity of judgements about evidence [21]. We can therefore expect non-experts in the domain to be less accurate in the assessment of the level of

Table 2 Concordance table that compares the experts' perceived level of evidence with the level of evidence from the literature, for each of the 95 indications (*n*)

Experts' perceived level of evidence[1]	Level of evidence from the literature[1]				
	RCT	Controlled trials, observational trials	Descriptive, experts' opinions	No proof	Marginal total [*n* (%)]
Reliable level of evidence	4	16	1	5	26 (27)
Weaker level of evidence	4	21	26	4	55 (58)
Experts' opinion	0	0	9	5	14 (15)
Own or peers' experience/opinion	0	0	0	0	0 (0)
Marginal total [*n* (%)]	8 (8)	37 (39)	36 (38)	14 (15)	95 (100)

RCT, randomized controlled trial.
[1]As defined in the Methods section.

Table 3 Relation between the experts' level of appropriateness and perceived level of evidence (*n*)

Experts' appropriateness categories	Experts' perceived level of evidence				
	Reliable level of evidence	Weaker level of evidence	Experts' opinion	Own or peers' experience/opinion	Marginal total [*n* (%)]
Appropriate	20	26	5	0	51 (54)
Uncertain	2	12	4	0	18 (19)
Inappropriate	4	17	5	0	26 (27)
Marginal total [*n* (%)]	26 (27)	55 (58)	14 (15)	0 (0)	95 (100)

Table 4 Concordance table comparing the level of appropriateness according to the literature and to the expert panel, for each of the 95 indications (*n*)

According to the panel of experts	According to the literature			
	Appropriate	Uncertain	Inappropriate	Marginal total [*n* (%)]
Appropriate	27	21	3	51 (54)
Uncertain	4	9	5	18 (19)
Inappropriate	4	14	8	26 (27)
Marginal total [*n* (%)]	35 (37)	44 (46)	16 (17)	95 (100)

evidence found in the literature [22]. The apparent difficulties of panel experts to appraise the level of published evidence underscores the need for a more explicit appraisal of the merits and limitations of the relevant studies in the literature review. For instance, a more systematic use of evidence tables and the GRADE approach may prove useful. In addition, during the panel meeting, the chair person possibly helped by methodologists fully aware of the relevant specific literature, should systematically include an explicit reminder about the strengths and limitations of existing evidence and a related discussion with the clinician experts, before they rate appropriateness criteria.

Compared with the 1994 panel of Swiss experts who rated appropriateness criteria for gastrointestinal endoscopy [11], appropriateness agreement between panel and literature was lower in our study. This is contrary to our initial hypothesis and unexplained. In fact, we would have imagined agreement to be higher when more homogeneous subsets of colonoscopy indications were considered, or when cases for which the balance between benefits and risks was easier to make, therefore leading to easier judgement of appropriateness. Differences in the way indications and/or studies were selected (more restrictive selection of articles in 1994 likely to lead to a higher concordance) may however be a possible explanation.

Lack of attention to the current literature, time lag between the publication of results and their knowledge, use in clinical practice, and lack of familiarity with the skills needed for the evaluation of evidence are some of the cited speculative reasons for lack of agreement between experts' and literature appropriateness [11]. In specific clinical situations, disagreement also appears when there is uncertainty regarding risks and benefits of a procedure. These reasons, among others, are considered to be weaknesses of RAM [23,24]. We also do not know which benefits and risks panellists were considering in their evaluations, and whether these were similar to those examined in the literature. In addition, other factors, such as clinicians' or patients' values and prior beliefs about treatment effectiveness, affect clinical decision making and may help explain why physicians may disagree on patients' management despite possible agreement on the evidence [25]. Our seven indications of full discordance between literature- and panel-based appropriateness categories were found to correspond, in an international multicentre observational study of the appropriateness of use of colonoscopy to 32 of 6004 colonoscopies (0.6%) [26]. A unique indication was encountered 28 times: high risk of colorectal cancer (non-polypomatous hereditary colorectal cancer) in an individual aged 20 years or more, with no previous colonoscopy. There were, however, no clear patterns that could explain the disagreements. In a few instances, literature-based evidence was relatively new and in opposition with a more usual practice reflected by experts' opinion.

The use of a multidisciplinary and international panel of experts, less prone to biases than single discipline or single country panels [18,27,28], the use of a validated method (RAM) and the conduct of sensitivity analysis were important strengths of this study. However, our study does have limitations. Firstly, the interpretation of concordance or of a kappa coefficient is impaired when the scales compared are not defined in the same way, or when the categories considered are not similar. Indeed, the categories of appropriateness and of levels of evidence we used for both the panel- and the literature-based approaches were not quite comparable. Therefore, the results can only be considered as indicative. In addition, categories were relatively broad and the uncertain category included indications for which no evidence was found in the literature. Secondly, the design only, and not the quality of the study, was considered. Misclassification of the level of evidence in the literature can therefore not be excluded. In addition, the modest agreement between the two persons who classified the level of evidence of the articles retrieved from the literature review indicates the actual difficulties in attributing a level of evidence to a specific study design and may thus contribute to understand the lack of concordance between the level of evidence from the literature and the level of evidence as perceived by the experts. Thirdly, the literature review was done in 1998, and several articles have been published since then. However, our objective was to emphasize the methodological issue of agreement and the exploration of concordance between two contemporaneous measures of level of evidence, rather than to specifically look at colonoscopy indications. Therefore, we would not expect major differences in our findings had the panel been organized more recently. Lastly, the limited number of experts (14) and indications (95) urges to cautious interpretation of some comparisons, because they were carried out on small numbers.

In conclusion, experts often overestimated the level of evidence available in the literature. In addition, only rarely did the experts' judgement completely disagree with the literature, although concordance between panel- and literature-based appropriateness was only fair. This underscores the need for panel-based methods of developing appropriateness criteria to include more explicit discussion about the strengths and limitations of existing evidence, with the clinician experts, before they rate appropriateness criteria.

Acknowledgements

The authors gratefully acknowledge the selfless commitment and invaluable contribution of all the expert panel members who made this project possible. This work was supported by the EU BIOMED II Programme (BMH4-CT96-1202), the Swiss National Science Foundation (32.40522.94 and 32.57244.99), and the Swiss Federal Office of Education and Science (95.0306–2).

References

1. Grimshaw JM, Thomas RE, MacLennan G *et al*. Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technol Assess* 2004; **8**: iii–iv, 1–72.
2. AGREE Collaboration. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Qual Saf Health Care* 2003; **12:** 18–23.
3. Grol R. Successes and failures in the implementation of evidence-based guidelines for clinical practice. *Med Care* 2001; **39:** II46–II54.
4. Atkins D, Eccles M, Flottorp S *et al*. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches. The GRADE Working Group. *BMC Health Serv Res* 2004; **4:** 38.
5. Burgers JS, Bailey JV, Klazinga NS, Van Der Bij AK, Grol R, Feder G. Inside guidelines: comparative analysis of recommendations and evidence in diabetes guidelines from 13 countries. *Diabetes Care* 2002; **25:** 1933–1939.
6. Hart RG, Bailey RD. An assessment of guidelines for prevention of ischemic stroke. *Neurology* 2002; **59:** 977–982.
7. Vogel N, Burnand B, Vial Y, Ruiz J, Paccaud F, Hohlfeld P. Screening for gestational diabetes: variations in guidelines. *Eur J Obstet Gynecol Reprod Biol* 2000; **91:** 29–36.
8. Thomson R, McElroy H, Sudlow M. Guidelines on anticoagulant treatment in atrial fibrillation in Great Britain: variation in content and implications for treatment. *Br Med J* 1998; **316:** 509–513.

9. Fahey TP, Peters TJ. What constitutes controlled hypertension? Patient based comparison of hypertension guidelines. *Br Med J* 1996; **313:** 93–96.
10. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff 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.
11. Nicollier-Fahrni A, Vader JP, Froehlich F, Gonvers JJ, Burnand B. Development of appropriateness criteria for colonoscopy: comparison between a standardized expert panel and an evidence-based medicine approach. *Int J Qual Health Care* 2003; **15:** 15–22.
12. Vader JP, Burnand B, Froehlich F, Dubois RW, Bochud M, Gonvers JJ. The European Panel on Appropriateness of Gastrointestinal Endoscopy (EPAGE): projects and methods. *Endoscopy* 1999; **31:** 572–578.
13. Froehlich F, Larequi-Lauber T, Gonvers JJ, Dubois RW, Burnand B, Vader JP. 11. Appropriateness of colonoscopy: inflammatory bowel disease. *Endoscopy* 1999; **31:** 647–653.
14. Bochud M, Burnand B, Froehlich F, Dubois RW, Vader JP, Gonvers JJ. 12. Appropriateness of colonoscopy: surveillance after polypectomy. *Endoscopy* 1999; **31:** 654–663.
15. Bochud M, Burnand B, Froehlich F, Dubois RW, Vader JP, Gonvers JJ. 13. Appropriateness of colonoscopy: surveillance after curative resection of colorectal cancer. *Endoscopy* 1999; **31:** 664–672.
16. Burnand B, Bochud M, Froehlich F, Dubois RW, Vader JP, Gonvers JJ. 14. Appropriateness of colonoscopy: screening for colorectal cancer in asymptomatic individuals. *Endoscopy* 1999; **31:** 673–683.
17. Guyatt GH, Haynes RB, Jaeschke RZ *et al.* Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. Evidence-Based Medicine Working Group. *JAMA* 2000; **284:** 1290–1296.
18. Koepsell TD, Weiss NS. *Epidemiologic Methods. Studying the Occurrence of Illness*. Oxford: Oxford University Press, 2003.
19. Vader JP, Froehlich F, Dubois RW *et al.* European Panel on the Appropriateness of Gastrointestinal Endoscopy (EPAGE): conclusion and WWW site. *Endoscopy* 1999; **31:** 687–694.
20. Atkins D, Best D, Briss PA *et al.* Grading quality of evidence and strength of recommendations. *Br Med J* 2004; **328:** 1490.
21. Atkins D, Briss PA, Eccles M *et al.* Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. *BMC Health Serv Res* 2005; **5:** 25.
22. Rosenbloom ST, Giuse NB, Jerome RN, Blackford JU. Providing evidence-based answers to complex clinical questions: evaluating the consistency of article selection. *Acad Med* 2005; **80:** 109–114.
23. Hicks NR. Some observations on attempts to measure appropriateness of care. *Br Med J* 1994; **309:** 730–733.
24. Phelps CE. The methodologic foundations of studies of the appropriateness of medical care. *N Engl J Med* 1993; **329:** 1241–1245.
25. Rubenfeld GD. Understanding why we agree on the evidence but disagree on the medicine. *Respir Care* 2001; **46:** 1442–1449.
26. Burnand B, Harris JK, Wietlisbach V *et al.* Use, appropriateness and diagnostic yield of screening colonoscopy for early colorectal cancer case-finding in asymptomatic patients: an international observational study (epage). *Gastrointest Endosc* (in press).
27. Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet* 2000; **355:** 103–106.
28. Fraser GM, Pilpel D, Kosecoff J, Brook RH. Effect of panel composition on appropriateness ratings. *Int J Qual Health Care* 1994; **6:** 251–255.

Accepted for publication 6 March 2006