Surg Endosc (2000) 14: 908–910
DOI: 10.1007/s004640000283

Surgical Endoscopy
Ultrasound and
Interventional Techniques

Consensus conferences must include a systematic search and categorization of the evidence

S. Sauerland, E. Neugebauer

Biochemical and Experimental Section, 2nd Department of Surgery, University of Cologne, Ostmerheimer Strasse 200, 51109 Cologne, Germany

Received: 16 November 1999/Accepted: 11 April 2000/Online publication: 29 August 2000

Abstract
Background: Ideally, a consensus panel combines expert knowledge with external evidence derived from the literature. To date, many consensus conferences do not use a structured approach to search the literature, but simply compile an add-on reference list from all papers cited by the panelists. This study examined how well such panelists retrieved the relevant literature.
Methods: We used the reference lists of nine surgeons who took part in a consensus conference on common bile duct stones. We included all papers that were referred to as randomized controlled trials (RCTs). We then compared this list with a database search in order to calculate sensitivity and specificity.
Results: The nine experts cited between 35 and 518 papers, but only eight papers on average were RCTs. Of the 49 papers that the experts believed to be RCTs, only 23 actually were RCTs. The sensitivity resp. specificity for correctly identifying an RCT was 0.21 (95% CI, 0.11–0.30) resp. 0.80 (95% CI; 0.64–0.95). RCTs that included the word "randomized" in their title were significantly more likely to be identified (relative risk, 1.31; 95% CI, 1.18–1.45).
Conclusion: Our data indicate that consensus panelists usually do not perform systematic literature searches, but simply use their favorite papers to back up their arguments. Because this may lead to a biased selection of the evidence base on which the consensus statements are founded, a systematic search of all relevant articles should become a mandatory task in any consensus or guideline process.

Key words: Clinical practice guidelines — Consensus development conference — Literature search — Publication bias — Retrieval bias

Correspondence to: E. Neugebauer

Because clinical practice guidelines (CPGs) aim at improving the process of health care and optimizing resource utilization, they have become increasingly popular during the last decade. They have been defined as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [6]. One of the most important techniques to systematically develop a clinical guideline is the formal consensus development process [7]. This technique was first adopted by the U.S. National Institutes of Health (NIH); subsequently, other institutions followed their lead [15, 18, 26].

Within the last few years, the advent of evidence-based medicine has partly changed the views of physicians on CPGs [3, 13]. Physicians who search the literature on their own and then base their clinical decisions preferentially on randomized controlled trials (RCTs) or systematic reviews of RCTs will accept and use CPG only if they distill the medical literature by appropriate methods [5]. Furthermore, the widespread development of CPGs by many different organizations has also lead to a wide spectrum of methodologies, some of which may be deficient. Not surprisingly, guidelines on the same topic have already reached different or even contradictory recommendations [8]. As the need for comparing guideline quality become apparent, several health care organizations throughout the world developed checklists to assess guideline quality quantitatively [11, 20, 22]. As a first result of these activities, a recent study by Shaneyfelt et al. [23] found that most CPGs use poor methodology, adding that the "greatest improvement is needed in the identification, evaluation, and synthesis of the scientific evidence."

In view of these considerations, the European Association for Endoscopic Surgery (E.A.E.S) has decided to critically evaluate the methodology of their consensus development conferences (CDCs). Since 1994, six CDCs on various surgical topics have been conjoined under the aegis of the E.A.E.S.. The relevant literature for these CDCs was searched by the panelists themselves. This study used the reference lists from a recent CDC on common bile duct

stones to evaluate how completely and how specificly these clinical experts were able to locate the relevant RCTs.

Methods

The consensus development conference

In 1996, the scientific committee of the E.A.E.S. decided to organize a consensus conference on common bile duct stones. The topic had been selected because the management of bile duct stones was undergoing major changes due to the advent of laparoscopic surgery. In January 1997, a group of experts was invited as panelists. The following criteria were used to select the panelist: (a) scientific and clinical expertise, (b) community influence, and (c) geographical location. We tried to create an interdisciplinary panel, inviting other experts as well as surgeons [19].

The panelists were asked to respond to eight specific questions relating to the etiology, diagnosis, and therapy of common bile duct stones. To support their responses, panelists were also asked to cite the relevant literature. We did not explicitly recommend the performance of a systematic literature seach. However, for one of the central questions, the experts were asked to use the AHCPR grading scheme [27] to categorize the literature according to its scientific validity.

In June 1997, a preconsensus statement was compiled out of the responses and sent out to all of the panelists together with the complete list of references, which were graded as class 1 or 2 evidence. During the annual conference of the E.A.E.S., the experts then met to discuss all of the questions in detail, modified the text of their recommendations, presented their results to the congress audience, and again modified the text. Because not all of the questions could be answered within the given time frame, a Delphi process was employed to finalize the CDC statements in October 1997. The full consensus statements were submitted to *Surgical Endoscopy* and published without revisions in 1998 [21].

Comparison of literature searches

In 1999, we again examined all of the reference lists sent by the panelists to the CDC organizers. All studies that were referred to as RCTs (evidence level 1) by the panelists were checked to see if they were truly randomized. We accepted all trials as RCT that reported being "randomized," regardless of whether or not the exact method of randomization was described. However, we compared the trials that used the word "random" in the title to those where it was mentioned only somewhere in the abstract or the methods section. We calculated the specificity of the panelists' literature search as the proportion of true RCTs among all studies that were referred to as RCTs by the panelists.

Next, we searched the literature by accessing the PubMed database with the Medical Subject Heading (MESH) "bile duct diseases" in combination with the Type of Publication (ToP) "randomized controlled trial." Only reports on common bile duct stones that were published before 1997 were considered. The reference lists of all panelists were then compared against this standard database of RCTs. To measure the sensitivity of the panelists' literature searches, we defined the retrieval rate as the number of relevant RCT identified by each of the panelists. All results are given with 95% confidence intervals (95% CI).

Results

Nine of the 13 CDC panelists graded their references according to scientific validity, so their responses could be used for the purposes of this study. These nine experts (eight surgeons and one gastroenterologist) cited an average of 124 scientific publications (range, 35–518), but only eight papers on average (range, 2–21) we referred to as RCTs. However, of the 49 papers that these experts believed to be RCTs, only 23 actually were RCTs. The specificity for identifying an RCT correctly was 0.80 (95% CI, 0.64–0.95).

The computer literature search resulted in a list of 84 RCTs. After we excluded the RCTs on biliary tract malignancies (n = 26), primary sclerosing cholangitis (n = 12), acute cholangitis (n = 9), and other biliary tract disease (n = 11), 26 papers on the treatment of common bile duct stones remained. If we consider this number to be the gold standard, the panelists' literature searches had an average sensitivity of only 0.21 (95% CI, 0.11–0.30). If we instead use only those RCTs as comparison that were cited by any other panel member, the sensitivity equals 0.23 (95% CI, 0.12–0.34). Two key papers were cited by six of the nine panelists, but none was cited by all of them. In three cases, RCTs were categorized as lower-level evidence (i.e., prospective cohort studies).

Interestingly, the only gastroenterologist on the panel cited two papers that nobody else in the group had located. RCT that included the word "randomized" in their title (n = 13) were significantly more likely to be identified by the panelists than trials that mentioned randomization only in the text (n = 10) (relative risk, 1.31; 95% CI, 1.18–1.45).

Discussion

It is widely suspected that "desk drawer" methods are inadequate for locating all of the relevant medical evidence that has been published on a specific medical question [2, 28]. Therefore recent approaches in research synthesis (e.g., the Cochrane Collaboration) have dedicated a great deal of work to formulating highly sensitive search strategies to locate all of the relevant RCTs [4, 17]. Until today however, the quality of the traditional approach for the identification of medical evidence has been examined quantitatively only for review articles [10, 12]. We used the example of a consensus-assisted guideline to assess the sensitivity and specificity by which consensus panelists retrieved RCTs from the literature.

Our results indicate that consensus panelists usually do not do a systematic literature search before participating in a consensus conference for guideline development. This reliance on the completeness of personal databases, however, leads to an incomplete awareness of the literature, even when the panelists are clinical experts. Because the primary literature searches had a sensitivity of only 20%, the completeness of the evidence was strongly dependent on the size and composition of the panel. Only the creation of a very large panel could ensure that all of the relevant articles have been identified by at least one panelist. An all-surgical panel is likely to overlook new technological advances described in nonsurgical journals, and vice versa. This reference bias may result in outdated or even misleading recommendations [9].

Apart from specialty, there are two other important barriers that prevent the complete incorporation of the medical literature in a CPG—language and publication type. Among the RCT reports that were considered in our CDC, there were two papers written in German and French and two abstracts. These publications were cited only by one panelist each, and one might expect that even more abstracts and non-English publications would have been located by handsearching the relevant journals and congress reports [5]. The use of other electronic literature retrieval systems, such as Embase, should also be considered to locate further relevant

material. Joyce et al. recently showed that authors of traditional review articles were more likely to cite papers that had been published in their specialty and in their home country [12].

Although only half of the RCT reports included the word "random" in their title—thus ignoring the CONSORT statement [1]—it does seem promising that the reports that were in accordance with the CONSORT statement were more likely to be identified. This finding lends support to the idea that standardized reporting of RCTs will increase the prompt identification and impact of rigorous research on clinical practice. Nevertheless, obviously, not all journals have adopted this concept yet.

A recent survey among French surgeons revealed that RCTs have only a very limited effect on clinical practice [16]. Three reasons were identified for this situation. The French surgeons read a mean of 40 issues of surgical journals per year, which may be too few to keep abreast with current medical advances. Furthermore, they strongly preferred journals published in French. Finally, only 59% of them could define the key characteristics of an RCT. In particular, older surgeons, who as experts are more likely to be invited to participate in developing CPGs, had only a very poor knowledge of clinical trial methodology. Thus, they attached more importance to other information sources—e.g., informal gatherings with colleagues and their own personal experience. This nonpublic dissemination of knowledge, however, can be looked upon as "back room politics," which hinder the implementation of CPGs [14]. As Sniderman so aptly described the problem, CPG development is often viewed as more of a social than a scientific affair [25].

In summary, we have provided strong data for the superiority of an evidence-based approach to guideline development. The lack of an entirely objective and complete method of identifying the medical evidence is probably the greatest weakness of traditional guidelines. In future, guideline developers who use a consensus technique for group processes should perform an independent search of the medical literature as a first step. Alternatively, a good synopsis of the literature can be achieved by summarizing the responses of a sufficiently large and multidisciplinary panel. However, we think it is important to emphasize that the recent advances in the methodology of systematic reviews should also play a role in the formulation of clinical guidelines [24]. This can probably be done most easily by first performing a systematic review and then using it as a basis for guideline formulation [29].

Acknowledgments. We are grateful to Dr. Antonius Helou, Hanover, Germany, for discussion and advice.

References

1. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz K, Simel D, Stroup DF (1996) Improving the quality of reporting randomiozed controlled trials. The CONSORT statement. JAMA 276: 637–639
2. Chalmers TC, Frank CS, Reitman D (1990) Minimizing the three stages of publications bias. JAMA 263: 1392–1395
3. Cook DJ, Greengold NL, Ellrodt AG, Weingurten SR (1997) The relation between systematic reviews and practice guidelines. Ann Intern Med 127: 210–216
4. Dickersin K, Scherer R, Lefebvre C (1994) Identification of relevant studies for systematic reviews. Br Med J 309: 1286–1291
5. Egger M, Davey Smith G (1998) Bias in location and selection of studies. Br Med J 316: 61–66
6. Field MJ, Lohr KN (1990) Clinical practice guidelines: directions of a new program. National Academy Press, Washington, DC
7. Fink A, Kosecoff J, Chassin M, Brook RH (1984) Consensus methods: characteristics and guidelines for use. Am J Public Health 74: 979–983
8. Fletcher SW, Fletcher RH (1998) Development of clinical guidelines. Lancet 352: 1876
9. Fraser GM, Pilpel D, Hollis S, Kosecoff J, Brook RH (1993) Indications for cholecystectomy: the results of a consensus panel approach. Qual Assur Health Care 5: 75–80
10. Gøtzsche PC (1987) Reference bias in reports of drug trials. Br Med J 295: 654–656
11. Hayward RS, Wilson MC, Tunis SR, Bass EB, Guyatt G, for the Evidence-based Medicine Working Group (1995) Users' guides to the medical literature. VIII: How to use clinical practice guidelines. A: Are the recommendations valid? JAMA 274: 570–574
12. Joyce J, Rabe-Hesketh S, Wessely S (1998) Reviewing the reviews: the example of chronic fatigue syndrome. JAMA 280: 264–266
13. Lohr KN, Leazer K, Mauskopf J (1998) Health policy issues and applications for evidence-based medicine and clinical practice guidelines. Health Policy 46: 1–19
14. Lomas J, Anderson G, Enkin M, Vayda E, Roberts R, MacKinnon B (1988) The role of evidence in the consensus process: results from a Canadian consensus exercise. JAMA 259: 3001–3005
15. McGlynn EA, Kosecoff J, Brook RH (1990) Format and conduct of consensus development conferences: multination comparison. Int J Technol Assess Health Care 6: 450–469
16. Millat B, Fingerhut A, Flamant Y, Hay JM, Fagniez PL, Farah A, Duron JJ, Courchevel JM, the French Association for Research in Surgery (1999) Survey of the impact of randomised clinical trials on surgical practice in France. Eur J Surg 165: 87–94
17. Muir Gray JA (1997) Evidence-based, locally-owned, patient-centred guideline development. Br J Surg 84: 1636–1637
18. Mullan F, Jacoby I (1985) The town meeting for technology: the maturation of consensus conferences. JAMA 254: 1068–1072
19. Neugebauer E, Troidl H (1995) Consensus methods as tools to assess medical technologies. Surg Endosc 9: 481–481
20. Ollenschläger G, Helou A, Kostovic-Cilic L, Perleth M, Raspe HH, Reinhoff O, Selbmann HK, Oesingmann U (1998) Checklist for methodological quality of guidelines: a contribution to quality promotion of medical guidelines [in German] Z Ärztl Fortbild Qualitätssich 92: 191–194
21. Paul A, Millat B, Holthausen U, Sauerland S, Neugebauer E, Scientific Committee of the European Association for Endoscopic Surgery (1998) Diagnosis and treatment of common bile duct stones (CBDS): results of a consensus development conference. Surg Endosc 12: 856–864
22. Petrie J, Barnwell E, Grimshaw J, for the Scottish Intercollegiate Guideline Network (SIGN) (1999) Criteria for appraisal of clinical guidelines for national use. http://www.show.scot.nhs.uk/sign/critmain.htm (accessed on June 9, 1999)
23. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J (1999) Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA 281: 1900–1905
24. Shekelle PG, Woolf SH, Eccles M, Grimshaw J (1999) Developing guidelines. Br Med J 318: 593–596
25. Sniderman AD (1999) Clinical trials, consensus conferences, and clinical practice. Lancet 354: 327–330
26. Stocking B, Jennett B, Spiby J (1991) Criteria for change: the history and impact of consensus development conferences in the UK. King's Fund Centre, London
27. U.S. Agency for Health Care Policy and Research (AHCPR) (1992) Acute pain management: operative or medical procedures and trauma. AHCPR, Rockville, MD
28. Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J (1999) Potential benefits, limitations, and harms of clinical guidelines. Br Med J 318: 527–530
29. Wortman PM, Smyth JM, Langenbrunner JC, Yeaton WH (1998) Consensus among experts and research synthesis: a comparison of methods. Int J Technol Assess Health Care 14: 109–122