Advance Access Publication 5 February 2006

eCAM 2006;3(1)3–12
doi:10.1093/ecam/nek017

Lecture Series

Evidence-Based Research in Complementary and Alternative Medicine II: The Process of Evidence-Based Research

Francesco Chiappelli[1,2,3], Paolo Prolo[1,2,3], Monica Rosenblum[4], Myeshia Edgerton[1,5] and Olivia S. Cajulis[6]

[1]Division of Oral Biology & Medicine, UCLA School of Dentistry, CHS 63-090, Los Angeles, [2]West Los Angeles Veterans Administration Medical Center, [3]Psychoneuroimmunology Group, Inc., [4]California State University, Northridge, [5]Tufts University Dental School and [6]Dental Group of Sherman Oaks, Inc., CA, USA

It is a common practice in contemporary medicine to follow stringently the scientific method in the process of validating efficacy and effectiveness of new or improved modes of treatment intervention. It follows that these complementary or alternative interventions must be validated by stringent research before they can be reliably integrated into Western medicine. The next decades will witness an increasing number of evidence-based research directed at establishing the best available evidence in complementary and alternative medicine (CAM). This second paper in this lecture series examines the process of evidence-based research (EBR) in the context of CAM. We outline the fundamental principles, process and relevance of EBR, and its implication to CAM. We underscore areas of future development in EBR. We note that the main problem of applying EBR to CAM at present has to do with the fact that the contribution of EBR can be significant only to the extent to which studies used in the process of EBR are of good quality. All too often CAM research is not of sufficient quality to warrant the generation of a consensus statement. EBR, nevertheless, can contribute to CAM by identifying current weaknesses of CAM research. We present a revised instrument to assess quality of the literature.

Keywords: evidence-based research – systematic review – consolidated standards of randomized trials – Markov model – complementary and alternative medicine

Evidence-Based Research

Aims and Caveats

Evidence-based research (EBR) in medicine, as conceived by A. Cochrane (1909–88), must not to be confused with medicine based on research evidence. EBR is a research movement in the medical sciences based upon the application of the scientific method. It seeks the conscientious, explicit and judicious identification, evaluation and use of the best evidence currently available. It is a systematic process whose purpose is to congeal the best available research findings with patient history and laboratory test results in order to optimize the process of making decisions about the care of each individual patient. Medicine based on the evidence, in contrast, is the traditional approach to medical treatment. It rests on long-established existing medical traditions, supplemented by individual pieces of evidence provided by the medical exam (e.g. history, test results), which may or may not have undergone adequate or sufficient scientific scrutiny (1–5).

The debate over evidence-based medicine versus medicine based on the evidence is complex, and far from being abated (4,6). It argues, for example, that medical doctors have depended upon reliable research evidence for their treatment ever since the rise of modern medicine (6). The EBR movement does not dispute that. It underscores the fact that research in the health sciences is advancing at such a fast pace that the body of evidence must be systematically

For reprints and all correspondence: Francesco Chiappelli, Ph D, Division of Oral Biology & Medicine, UCLA School of Dentistry, CHS 63-090, Los Angeles, CA 90095-1668, USA. Tel: +1-310-794-6625; Fax: +1-310-794-7109; E-mail: chiappelli@dent.ucla.edu

evaluated and synthesized for benefit of patients, providers and society (4,7,8).

A second argument stems from the fact that, in some domains of the health sciences at least, the research evidence can be deficient, inadequate or unreliable, and that therefore medicine must rest on traditional modes of interventions even if they have not been validated by research (6). The EBR movement underscores in this context that it is through the systematic evaluation of the research methodology, designs and data analysis that it becomes possible to identify research deficiencies in given clinical domains, which then serve to improve quality of research evidence (4,5,7,9).

A third important point of argument suggests that proponents of EBR make a conceptual error by grouping knowledge derived from clinical experience and physiological rationale under the heading of the best available evidence, and further compound errors by developing hierarchies of evidence. That is to say, lack of evidence and lack of benefit are not the same, and the more data are pooled and aggregated the more difficult it becomes to compare patients in studies with the individual patient in front of the doctor. Clinicians need to incorporate knowledge from several distinct areas into medical decision, including empirical evidence, experience, physiological principles, patient needs, wants and coverage, and professional values (6). This latter question is particularly relevant to an unbiased appreciation of EBR, and the remainder of this paper responds to this question, with emphasis on complementary and alternative medicine (CAM).

Consensus of the Best Available Evidence

Certain caveats plague the practical application of EBR in a day-to-day medical practice (2–4,10), particularly in the context of certain CAM protocols, such as acupuncture (9). A few salient among these are listed in Table 1. It is also true, however, that the fundamental purpose of EBR is to validate modern medical practice, and consequently the evolution and establishment of evidence-based medical practice is a *sine qua non* for medicine in the 21st Century (7).

EBR contributes to the validation of medical practice by systematically evaluating strength of available evidence (2,4,5,7).

Table 1. Fundamental limitations of EBR

- Overwhelming scope of the scientific information.
- High stringency of scientific research.
- Challenge to maintain up-dated research evaluation.
- Demands of clinical relevance versus statistical significance.
- Different views on clinical relevance (e.g. levels of clinical significance based on categories, such as tangible versus intangible benefits, size of treatment effect).
- Subjectivity in evaluation of internal versus external validity threats.
- Lack of clinical use and acceptability through clinical testing.
- Guarded stance at the prospect of changing and amend intervention protocols.

The purpose of EBR is not to group knowledge derived from clinical experience and physiological rationale under the heading of best available evidence (6), nor is it to develop hierarchies of evidence (6). Quite the contrary, EBR aims at generating a consensus statement that summarizes the outcome of a process of systematic evaluation of the literature. The statement provides *ipso facto* scientific validation of the best available evidence thus generated from all of the available research and of the clinical decision-making process (2,7,8,10).

The consensus statement is the outcome of the process of the systematic review and evaluation of all of the available evidence. It presents inferences, summative evaluations and conclusive narrative synthesis of the findings. It discusses problems pertaining to presentation and relevance of findings, including whether or not key elements of each study are clearly displayed, magnitude of findings is statistically significant and the findings are homogeneous or heterogeneous. The consensus statement also addresses concerns of clinical relevance, of the validity of the integration process (e.g. inclusion and exclusion criteria, comprehensive search strategy) and of the rigor of the evaluation process (e.g. quality of evidence rating, cf. double arrow in Fig. 1). The focus of the consensus statement pertains to sensitivity and specificity analyses, and whether or not the overall findings suggest an overall net benefit for patients. To assess the quality control of the process of integration, a third independent reviewer, 'standardized' to the other readers (11), usually is engaged to assess systematically the studies' validity and statistical and clinical significance. The consensus statement includes a discussion of those issues as well. In brief, the consensus statement discusses the quality of the evidence on each individual report, as well as a bottom-line statement, a cogent synthesis of the research, explicating the best available evidence (2–5,7).

The panel of experts who performs the systematic review drafts the consensus statement. It is then presented and discussed in an open forum to patient group advocates and the general public. The panel finalizes the consensus statement in executive session, and the final report is generated. Some

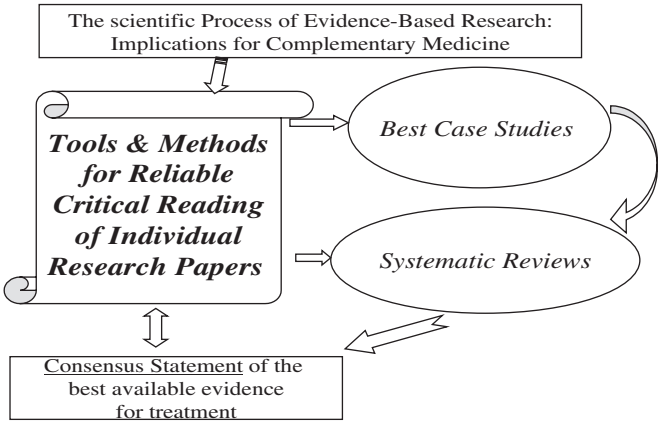


Figure 1. The process of evidence-based research in complementary and alternative medicine.

review groups (e.g. the Cochrane Group) insist upon the need for regular updates of the consensus statement (usually every 6 months), some others do not (e.g. the National Institutes of Health, NIH).

Developing the Consensus Statement

The overall report generated at the completion of the EBR process evidently goes well beyond the routine narrative literature review. It is a systematic review of all of the available research evidence—good and bad (on criteria of research design, methods and data analysis)—which culminates in the consensus statement. The systematic process of critical evaluative research of the available evidence follows the scientific method (2,5,7). It is not aimed at pooling and aggregating data across studies indiscriminately. EBR aims at determining the quality of each report, based on stringent criteria of research methodology, design and data analysis. Reports deemed acceptable are used in the second phase of the process that synthesizes the evidence by means of meta-analytical techniques, and generates a bottom-line, which serves to aid clinical decision-making (1,12).

The goal is clearly not to compare the patients in the studies with the individual patient in front of the doctor (6). It is to apply best of what research literature provides for direct benefit of patients in front of doctor (4,5,10).

Clinical research in CAM in the 21st Century requires the stringent, rigorous and systematic approach provided by EBR. The paucity of CAM specific validated measures or use of more generic measures will impact directly on the EBR process: the significant current debate around what outcomes should be measured and how they are measured is not abated (13). The future of clinical and translational research in CAM lies in the systematic evaluation of research evidence in treatment intervention for patients and in judicious and timely generation of the consensus statement (7,9,14,15).

Research on Research

The Process of EBR

The flow of EBR, outlined in Fig. 1, is applied to the performance of systematic reviews, which encompass all available literature. Best case studies in EBR entail performance of the process of EBR with a random sample of available literature. The scientific process of EBR is dependent upon essential tools and methods for the reliable qualitative and quantitative critical reading of individual papers in the context of the best-case studies and, more broadly, of systematic reviews. This figure illustrates that the end-product, the bottom-line of EBR, is the generation of a consensus statement, as discussed above. This figure also indicates a reciprocal feedback between the box of tools for EBR and the consensus statement. This is so because the complete process of EBR employs a set of selected and specified tools and instruments of research, which generates the analysis of findings, which is presented and discussed in the consensus statement. The consensus statement thus should ideally include a discussion of strengths, weaknesses, limitations and caveats of these tools and instruments. The double arrow is meant to represent this reciprocal feedback by which the generation of the consensus statement is derived from the use of certain tools of research, and provides an evaluative component with respect to if and how these tools ought to be perfected for future evidence-based research.

EBR is a Form of Critical Research on Research that Follows the 5-Step Scientific Process

(i) It begins by stating the research question, which comprises the PIC/PO question. The question defines the patient population being examined and the interventions being considered (e.g. conventional treatment versus conventional treatment supplemented with CAM), whether the interventions are compared or studied from the longitudinal perspective and predictions are being drawn, and specifies the outcome of interest (5,7).

(ii) The second step involves methodological issues, including the sampling and accessing of the research literature, and the tools for critical analysis of the reports (5,7). The sampling process requires extensive library search of published materials (e.g. clinical trials) and additional individual communications with individual researchers and authors, when further information is needed.

(iii) The sample is critically evaluated using stringent standards [e.g. the Consolidated Standards of Randomized Trials (CONSORT) (16,17)]. In the case of acupuncture, the STRICTA norms (Standards for Reporting Interventions in Controlled Trials of Acupuncture) are further recommended (9). Reliable and valid instruments {e.g. Timmer scale, Jadad scale, Wong [cf. Appendix 1; (18)], Linde internal validity scales; for a review, see (5)} are used for this purpose. Alternative means [e.g. GRADE, ASSERT; for a review, see (19)] are also utilized and converge with the former in quantifying levels of quality of the research and of levels of significance of evidence.

(iv) The data from separate reports are pooled, analyzed for acceptability (20), and when appropriate, utilized in meta-analysis or meta-regression analyses for the generation of an overarching statistical significance (5,12,21,22). EBR data can also be analyzed by Individual Patient Data (IPD) (23) or Number Needed to Treat (NNT) analyses (3). These formats differ from traditional modes of statistical analysis in that they pertain to analysis of data of individual patients as opposed to traditional analysis of data from groups of patients. In general, moreover, EBR data are best

analyzed by means of Bayesian, rather than the traditional Fisherian statistics, in order to interpret data from research in the context of statistical significance and clinical relevance.

(v) The last step is a cumulative synthesis, which summarizes the process and the findings. The consensus statement must be coherent with and reflect the best available evidence with respect to the stated PIC/PO question [cf. Appendix 2 adapted from (7)] (2,10,24,25).

Merits and Strengths of EBR in Clinical Decision-Making

The merit and strength of EBR lie not only in the rigor of its scientific method but also in the validity of its product, the consensus statement. EBR and the outcomes it generates have direct applications and extensions to immediate needs of patients, to the best available evidence for intervention and cost (2,7,8,10). A well-constructed consensus statement presents a cost-effectiveness analysis, which is a process of decision analysis that incorporates risks as well as cost. This is achieved by a step approach that generally involves an evaluation of whether or not the problem was framed in a clinically relevant manner (i.e. PIC/PO question), of the validity of integrated information (i.e. critical evaluation of the literature), the rigor of process of integration [i.e. inclusion & exclusion criteria of reliable versus unreliable (acceptable versus unacceptable) evidence] and of the presentation and quality of the findings (i.e. summative evaluation) (7).

The relevant findings in this cost-effectiveness analysis are most often expressed as the incremental cost-effectiveness between conventional treatment alone and conventional treatment supplemented by complementary alternative treatments. The incremental ratio, that is the difference in costs between the two strategies divided by the difference in effectiveness between the two strategies, is often presented as well (Fig. 2) (5).

Figure 2 illustrates that following the scientific process of EBR and generation of the consensus statement, the overall clinical relevance is assessed, implemented and evaluated by the clinician. Effectiveness and utilities data are estimated (e.g. Markov model; cf. Appendix 3: the Markov Process) to aid the final clinical decision-making process. (5,10).

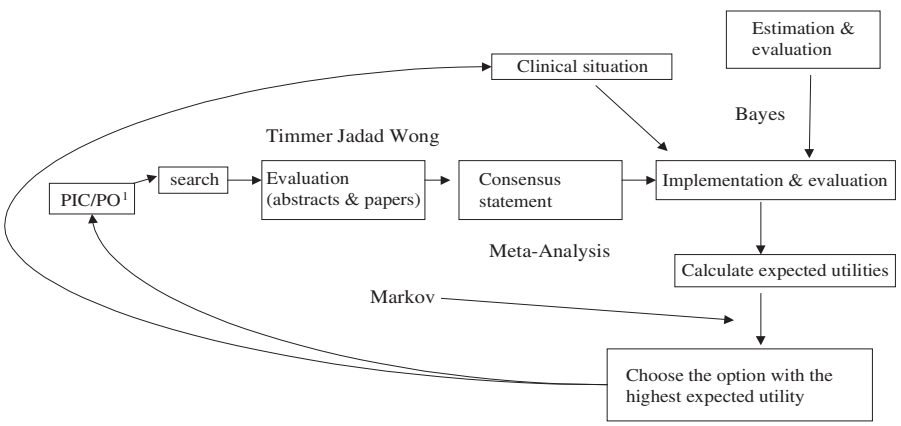
The EBR process evaluates each competitive strategy, usually by means of the Markov model-based decision tree. This approach permits to model events that may occur in the future as a direct effect of treatment or as a side effect. The model produces a decision tree that cycles over fixed intervals in time, and incorporates probabilities of occurrence. Even if the difference between the two treatment strategies appears quantitatively small, the Markov model outcome reflects the optimal clinical decision, because it is based on the best possible values for probabilities and utilities incorporated in the tree. The outcome produced from the Markov decision analysis is generally obtained by means of the sensitivity analysis to test the stability over a range probability estimates, and thus reflects the most rational treatment choice (25,26).

EBR in the Context of CAM

In summary, the performance of EBR is a science in its own right. The integration of the EBR paradigm in CAM has already been recognized (7,14,15,27). Undoubtedly, EBR will increasingly play an important role in distinguishing appropriate versus non-appropriate (i.e. acceptable versus non-acceptable; see below) CAM-based intervention in the future. Clinical and translational CAM research in the 21st century will rely upon the systematic evaluation of the research evidence. Progress in EBR of CAM must strive along these dimensions.

Tools and Protocols

First, the field of EBR needs to refine and finalize its tools and protocols. The critical process in EBR entails the critical



[1]PIC/PO: problem/population – intervention – comparison/prediction – outcome

Figure 2. Algorithm of the process of applying research evidence in clinical decision making [adapted from (5)].

evaluation of the research methodology, design and data analysis. Depending upon the tools utilized to evaluate the scientific literature, scores are obtained about the completeness and quality of research methodology, and design and statistical handling of the findings are generated (SESTA, systematic evaluation of the statistical analysis). Appendix 1 offers a revision of the Wong scale (18), followed by a detailed highlight of the SESTA paradigm. Utilization of this scale and of SESTA permit the qualitative and quantitative evaluation of the research methodology, design and data analysis. Quantification yields values that are analyzed by acceptable sampling statistical protocols to establish whether or not the sample of research reports studied by means of the evidence-based process has met criteria of acceptability to produce meta-analyses and reliable over-arching inferences.

This protocol does not seek to estimate the quality of the lot, which would be equivalent to evaluating the quality of the search literature process, but rather to estimate its acceptability (20). Acceptance sampling generates information based either on the attributes (i.e. nominal variable: acceptable based on a set of rigorously set criteria versus not acceptable) or on the characteristics of the identified information (i.e. continuous variable assessed along some interval scale).

Case in point, a best-case study we conducted of the use of music therapy as an alternative intervention to relieve anxiety generated reliable data, which permitted demonstration of the relative consistencies and inconsistencies in research methodology, research design and data analysis across the papers evaluated in the systematic review. The data could be used to quantify and to underscore strengths and deficiencies of this specific domain of the CAM literature. This best case study on music therapy as an alternative mode of intervention for anxiety revealed that the two weakest domains of that research literature pertain to information provided about the number needed to treat and statistical analysis of data. Another overwhelming weakness of this literature relates to the tools of measurement. Acceptable sampling analysis of these findings indicated that these three deficiencies were statistically significant (Greenhouse-Geisser $F = 7.58$, $P < 0.0001$; Scheffé, $P < 0.05$). Of borderline significance ($P < 0.1$) was the domain of research that pertains to the establishment of statistical and clinical significance.

In brief, this analysis permits to evaluate strength and stringency of music therapy as an alternative mode of intervention for anxiety. It established that 90% of this literature has appropriate research methods, design and data analysis, with an overall score (21.09 ± 3.14) within the 95% confidence interval of top rating. It also identified the principal domains of weakness within this CAM literature (e.g. number needed to treat information, statistical analysis of the data, tools of measurements), which must be corrected in future research. Lastly, this analysis underscored the fact that the literature on music therapy for anxiety has to date failed to make a compelling statement of relationship between statistical significance of findings and their clinical relevance.

Future Analyses in EBR

For the future, it is important to realize that these research quality-rating scales lead to the possibility of an evaluation following the principles of Boolean logic. That is to say, if, for instance, the first two questions (i.e. study question and study outcome) are evaluated to be congruent, then a conjunctive logic association is produced (study question = 1, study outcome = 1, conjunction = 1). This outcome then leads to evaluation of whether or not the measures and design are in fact congruent with the study question and outcome as reported. A conjunctive logic association furthers the process to examine SESTA, which itself can be reduced to a series of Boolean arguments. The outcome of the process, which we are now in the process of automating in a computer-assisted software, is either 1 (report overall acceptable based on criteria of research methods, design and analysis) or 0 (report unacceptable). Zeros appear in the Boolean process whenever a disjunction is attained (e.g. design in congruent for stated study question and study outcome). Both acceptable and unacceptable reports are integrated into the consensus statement. The latter contributes in formulating recommendations for the best available evidence for clinical decision-making, whereas unacceptable reports are discussed in terms of their deficiencies and the information they may provide for further improvement of research.

Evaluations of EBR such as these hold considerable promise to strengthen quantification of the EBR process, and thus to enhance considerably the value of the consensus statement. This will contribute to the role and valence of EBR in providing informed and scientifically supported statements of acceptability for CAM.

Practicality and Dissemination of EBR in CAM

The dissemination of EBR in CAM must become more practical and contextual, in order for it to become intelligible to providers, to patient groups and to insurance carriers. This is necessary to facilitate its integration into every-day medical decision-making and treatment (5,10,28). This will require concerted efforts to expand and to deepen education about knowledge of the process, outcome and practical uses of EBR (24), and to utilize EBR in daily procedures and protocols in order to shift from 'trade-professions' to 'evidence-based professions'. Existing meta-analyses and systematic reviews (e.g. Cochrane reports) should be catalogued, reviewed and summarized, and their findings should be effectively disseminated among providers, patients and insurance providers (17).

EBR Specialist for Benefiting CAM Practice

Lastly, the establishment of an 'EBR specialist' must be seriously considered (3), who can work from within the medical establishment to retrieve, read, evaluate and present the best available evidence with respect to complementary and alternative modes of intervention. This specialist will contribute to the establishment of criteria of EBR and of evidence-based

clinical practice guidelines that will require these to be validated, assessed and monitored by a network of professional EBR practitioners in CAM under the auspices of national and international professional, medical and CAM associations.

The EBR specialist will contribute to the process of review, assessment and evaluation of consensus statements, as well as of complaint for malpractice based on evidence-based clinical practice guidelines. This latter point is particularly important in the context of CAM interventions, which are often prone to distrust by the Western medical establishment because of the lack of substantiating research evidence. The EBR specialist will endeavor, for instance, at disseminating findings of systematic reviews on CAM through the Internet to make EBR easy to access, easy to understand and easy to use. This will require dissemination of consensus statements in lay and foreign languages.

In summary, the concerted efforts we have outlined hold the promise of increasing acceptance and dissemination of CAM treatment modalities for the ultimate benefit of patients (29,30). This endeavor can only be attained by stringent adherence to the scientific method in EBR over the next decades (7). EBR can be a powerful tool for identifying the questions for which no satisfactory evidence exists—a very common situation in CAM. Nevertheless, it must also be recognized that the main problem of applying EBR to CAM has to do with the fact that EBR will be useful in this context of science only to the extent to which the studies used in the process of EBR are of good quality, comparable and reliable. Unfortunately, all too often studies on CAM modalities are still today of inferior quality and preclude a sound EBR approach. Acceptable analysis of the type described above, and the use of certain EBR instruments, will serve to identify the weaknesses of CAM research.

Acknowledgements

The author thanks the students and colleagues of the UCLA Evidence-based Research Group for their contribution. The author is indebted to Dr Michael Newman, Dr Negoita Neagos, Dr Javier Iribarren and Dr Janet Bauer for the discussions leading to this work. This study was supported in part by funds from the UCLA School of Dentistry and the Alzheimer's Association.

References

1. Greenhalgh T. How to read a paper: papers that summarise other papers (systematic reviews and meta-analyses). *BMJ* 1997;315:672–5.
2. Friedland DJ, Go AS, Davoren JB, Shlipak MG, Bent SW, Subak LL, Mendelson T. *Evidence-based Medicine: A Framework for Clinical Practice.* Stamford: Appleton & Lange, 1998.
3. Chiappelli F, Prolo P. Evidence-based dentistry for the 21st century. *Gen Dent* 2002;50:270–3.
4. Abt E. Complexities of am evidence-based clinical practice. *J Evid Based Dent Pract* 2004;4:200–9.
5. Chiappelli F, Prolo P, Negoatis N, Lee A, Milkus V, Bedair D, Delgodei S, Concepcion E, Crowe J, Termeie D, Webster R. Tools and methods for evidence-based research in dental practice: preparing the future. In: *Proceedings of 1st Int Conf Evidence-Based Dental Practice*. *J Evidence Based Dental Pract* 2004;4:16–23.
6. Tonelli MR. The limits of evidence-based medicine. *Respir Care* 2001;46: 1435–40.
7. Manheimer E, Berman B. Cochrane for CAM providers: evidence for action. *Alt Therap Health Med* 2003;9:110–2.
8. Prolo P, Weiss D, Edwards W, Chiappelli F. Appraising the evidence and applying it to make wiser decisions. *Braz J Oral Scien* 2003;2:200–3.
9. Hammerschlag R. Acupuncture: on what should its evidence be based. *Altern Ther Health Med* 2003;9:34–5.
10. Bauer J, Spackman S, Chiappelli F, Prolo P. Model of Evidence-Based Dental Decision-Making. *J Evid Based Dent Pract* 2005 (In Press).
11. Ramos KD, Schafer S, Tracz SM. Validation of the Fresno test of competence in evidence-based medicine. *BMJ* 2003;326:319–21.
12. Thorton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol* 2000;53:207–16.
13. Mason S, Tovey P, Long AF. Evaluating complementary medicine: methodological challenges of randomised controlled trials. *BMJ* 2002;325: 832–4.
14. Chang I-M. Initiative for developing evidence-based standardization of Traditional Chinese medical therapy in the Western Pacific region of the World Health Organization. *eCAM* 2001;1:337–42.
15. Walach H, Jonas WB, Lewith. The role of outcomes research in evaluating complementary and alternative medicine. *Altern Ther Health Med* 2002;8: 88–97.
16. Moher D, Schultz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134:657–62.
17. Moher D, Soeken K, Sampson M, Ben-Porat L, Berman B. Assessing the quality of reports of systematic reviews in pediatric complementary and alternative medicine. *BMC Pediatr* 2002;2:3–10.
18. Wong J, Prolo P, Chiappelli F. Extending evidence-based dentistry beyond clinical trials: implications for materials research in endodontics. *Braz J Oral Sci* 2003;2:227–31.
19. GRADE Working Group. Grading quality of evidence and strength of recommendation. *BMJ* 2004;328:1–8.
20. Montgomery DC. *Introduction to Statistical Quality Control*, 4th edition. New York: Wiley & Sons, 2000.
21. Lau J. Schmid CH, Chalmers TC. Cummulative meta-analysis of clinical trials builds evidence for exemplary medical care. *J Clin Epidemiol* 1995;48:45–57.
22. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA* 2000;283:2008–12.
23. Nieri M, Clauser C, Pagliaro U, PiniPrato G. Individual patient data: a criterion in grading articles dealing with therapy outcomes. *J Evid Base Dent* 2004;3:122–6.
24. Carney PA, Nierenberg DW, Pipas CF, Brooks WB, Stukel TA, Keller AM. Educational epidemiology: applying population-based design and analytic approaches to study medical education. *JAMA* 2004;292: 1044–50.
25. Sugar CA, James GM, Lenert LA, Rosenheck RA. Discrete state analysis for interpretation of data from clinical trials. *Med Care* 2004;42:183–96.
26. Yu F, Morgenstern H, Hurwitz E, Berlin TR. Use of a Markov transition model to analyse longitudinal low-back pain data. *Stat Methods Med Res* 2003;12:321–31.
27. Teresawa K. Evidence-based reconstruction of Kampo medicine: Part-III-How should Kampo be evaluated? *eCAM* 2004;1:219–22.
28. Prolo P, Weiss D, Edwards W, Chiappelli F. Appraising the evidence and applying it to make wiser decisions. *Brazilian Journal of Oral Sciences* 2003;2:200–3
29. Hankey A. CAM modalities can stimulate advances in theoretical biology. *eCAM* 2005;2:5–12.
30. Rangel JAO. The systemic theory of living systems and relevance to CAM. *eCAM* 2005;2:13–8.

Received August 3, 2005; accepted January 5, 2006

Appendix 1: Wong Scale (Revised)

1. What

A. What is the research question/purpose/outcome sought? Is the stated purpose tested and measured correctly?

2–3 sentences

Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

B. What are the findings, how are they presented? Do the findings respond to the stated purpose/outcome sought?
2–3 sentences
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

C. What is the clinical significance of the findings, and what is their statistical significance? Do the findings mean anything anyway . . . research-wise or clinic-wise?
2–3 sentences
Note: issues about risk-to-benefit ratio, cost-to-benefit ratio; also note issues about P-values versus α level
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

2. Who
A. What was the sample tested, is the sample representative of the population under study, of your patients?
2–3 sentences
Note: issues of sampling and related threats to external validity (i.e. selection of sample representative of population under study) and to internal validity [i.e. concerns of maturation, mortality (i.e. drop-out), history]. Note that this question includes the drop-out query of the Jadad scale.
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

B. Are numbers presented in the paper that you can trust, and would that permit you to compute the Number Needed to Treat (NNT)? List experimental group event rate (EER) and control group event rate (CER), and compute NNT. Is there any information about Intention to Treat (ITT)
2–3 sentences
Note: this question pertains ONLY to Clinical Trials Outline ITT, if information is provided
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

C. Can the information provided in the paper be of any use directly to any patient or the group of patients in your practice, now?
2–3 sentences
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

3. How
A. How was the question addressed from the perspective of design, and were the appropriate caveats discussed?
2–3 Sentences
Note: distinguish between Prognostic/Diagnostic studies, between observational (prospective, cross-sectional, case-control) and experimental designs, and within these clinical trials (run-in, cross-over, mixed). Note issues of randomization and blinding (related to two of the Jadad queries). Note issues of correct selection of the control and experimental groups (threats to internal validity)
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

B. How was the outcome measured, were issues of reliability and validity presented?
2–3 sentences
Note: issues of selection of the instrument (threat of internal validity) to measure the outcome variable under study; issue of reliability (inter-rater, intra-rater, internal consistency) and validity (criterion, content, construct) of measurement.
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine

C. How were the data presented and analyzed (SESTA)?
2–3 sentences
Note: refer to fundamental elements of SESTA below
Score: 1 = inadequate, raise doubts and concern, 2 = adequate, but incomplete, 3 = fine
Note: Fundamental Elements of SESTA

What is the analysis meant to do?
Categorical versus Continuous Data; Comparison versus Prediction; time series versus survival

Categorical Data Analysis

- Are categorical data, and only categorical data, analyzed by the χ^2–test? (y/n):
- Are the data matched categorical data and is the ***McNemar*** χ^2-test used? (y/n):
- Are the categorical data analyzed as a difference from baseline, and therefore the ***Cochran Q*** χ^2-test utilized? (y/n):
- Are the categorical data analyzed from the perspective of a prognostic stratifying variable, and is the ***Mantel-Haenzel*** χ^2-used? (y/n):
- Is the research design a 2×2 format and therefore uses the ***Yates Correction for Continuity***? (y/n):
- Are the E values (expected frequencies) >5, and therefore the ***Fisher's Exact test*** used? (y/n):
- Is collapsing performed judiciously, and does it involve contiguous cells? (y/n):
- Is care given to avoid overly strong and absolute conclusions, which the weak nature of the χ^2-test generally does not warrant? (y/n):

Continuous Data Analysis—Comparison

- Are the ***assumptions for parametric statistics*** (i.e. normality, independence, homogeneity of variance) mentioned, tested, not violated? (y/n):
- If the assumptions are satisfied, are the data correctly analyzed by a ***matched t-test***? (y/n):
- If the assumptions are not satisfied, are the data correctly analyzed by the ***Signed-Rank Wilcoxon/Mann–Whittney U-test***? (y/n):
- In the case of a comparison of only two groups with no matching, if the assumptions are satisfied, are the data correctly analyzed by a ***Student t-test***? (y/n):

- If multiple outcome variables are compared in two groups, is the ***Hötelling T^2-test*** presented? (y/n):
- If the assumptions are not satisfied, are the data correctly analyzed by the ***Rank Sum Wilcoxon/Mann–Whitney U-test***? (y/n)
- If the assumptions are satisfied, and more than two groups are studied, are the data correctly analyzed by the ***analysis of variance (ANOVA)***? (y/n):
- If the assumptions are not satisfied, then is a ***Geisser Greenhouse correction*** presented? (y/n):
- Is one of the control variables used as a covariate—that is, a variable that can vary together with the outcome measure—and which ought to be used 'to correct measurements of the outcome measure in order to obtain the true and correct outcomes to be analyzed? (y/n):
- If the assumptions are satisfied, are the data correctly analyzed by the ***analysis of covariance (ANCOVA)***? (y/n):
- If ANOVA or ANCOVA were performed, are ANOVA or ANCOVA tables complete with sums of squares, degrees of freedom, mean squares, F-values and *P*-values presented? (y/n):
- **OR** are F statements not presented in the text with, in brackets, the Fcrit value, degrees of freedom, and *P*-value? [e.g. F(3.45), df = 5; $P = 0.001$] (y/n):
- **OR** are *P*-values not simply scattered in the text? (y/n):
- If a significant F-value is established by ANOVA or by ANCOVA, are main effects and interactions tested by ***pre hoc or by post hoc comparisons***? (y/n):
- Was care taken in correcting the α level for the number of repeated comparisons within the design by means of the ***Bonferroni Correction***—or any type of correction? (y/n):
- Upon performing *post hoc* comparisons, and drawing conclusions from them, was care taken of using the appropriate test (e.g. comparing all the group means in the design by the ***Scheffe's test***, comparing all possible pairs of means of ***Tukey's Honestly Significant Difference test***, comparing pairs of means following a ranking process by using ***Newman-Keul's test***, comparing means in a stepwise fashion to a reference control group with *Dunnett's* test)? (y/n):
- If the design involves more than two groups, but any one of the three assumptions are not satisfied, does the design involve one independent variable—'one way' and is correctly analyzed by the ***Kruskal–Wallis test*** ? (y/n):
- **OR** does it involve two or more independent variables—or one independent variable and one or more control variables, 'factorial' design—and is correctly analyzed by the ***Friedman test***? (y/n):

Continuous Data Analysis—Association and Prediction

- If the data presenting associations between two variables, is the ***Pearson Correlation Coefficient*** correctly used only when both variables are continuous? (y/n):
- Is the ***Spearman Rho correlation coefficient*** correctly used when one of the two variables is categorical? (y/n):
- Is the Cohen κ ***coefficient*** correctly computed and discussed in the instance of agreement between two observers along a binary—diseased/not diseased—variable? (y/n):
- Are causal relationships not erroneously drawn from correlations? (y/n):
- If the data is presented in a prediction model, are standardized ***regression coefficients*** shown as ***beta weights*** and their statistical significance established? (y/n):
- Is the significance of the predictive model established by means of ***ANOVA analysis***? (y/n):
- Is the overall relationship among the predictors established by the ***R^2 estimate***? (y/n):
- In establishing the hierarchical predictive strength of each predictor, is hierarchical or a stepwise regression model adopted? (y/n):
- In the case of a binary—diseased/not diseased—non-continuous outcome measure, is the ***logistic regression model*** utilized? (y/n):
- Is the ***goodness-of-fit***, the difference between observed and fitted probabilities, presented and discussed? (y/n):

Time-Series and Survival Data

- If the data present time-related analyses, are these analyses short-term, and are the data analyzed correctly in a ANOVA design that involves ***repeated measures***? (y/n):
- OR, do the data relate to long-term time-series design, and are presented by means of life tables, which are analyzed by means of the ***Kaplan–Meier***, and the ***Cox proportional hazard*** regression analysis? (y/n):

Overall Evaluation

2–3 paragraphs to highlight the strengths and weaknesses of the paper and defend the overall critical evaluation and score.

Score

Each question is scored from 1 to 3. The total WWH score ranges from 9–27 for Clinical Trials, and from 8–24 for studies that are not Clinical Trials, and where the NNT does not apply

Appendix 2: EBR Recommendations in CAM

[adapted from (7)]

I. Clinical Relevance

(i) Determine whether or not a systematic review is relevant to patient care

(ii) Establish a clearly defined and clinically relevant research question expressed in terms of the relation between a CAM test intervention and a control comparison

II. Sampling Criteria

(i) Explicitly define the inclusion and exclusion criteria appropriate for identifying the studies used to answer the clinical CAM question

(ii) Criteria optimally include the following:

randomized controlled trials (RCTs) and quais-RCTs
patient groups must pertain to the patient population under study
interventions to be compared must relate to the study question
studies identified in the search process must present assessments of the outcome measure under investigation

(iii) Establish a systematic search strategy for comprehensive sampling of available studies, which must include foreign and non-customary bibliographic databases [whether the inclusion of 'gray' literature (i.e. non-peer-reviewed, public domain) is recommendable or not is debatable because of potentially unwise investment of resources]

III. Quality of the Evidence

(i) Characterize the threats to internal (i.e. replicability) and to external validity of the study (i.e. generalizability) by means of research quality-rating scales (e.g. Jadad, Wong, Timmer, Linde, GRADE)

(ii) Extract and tabulate data pertinent to meta-analysis (e.g. group sample sizes, and means and standard deviations of outcome under study)

(iii) Establish feasibility of meta-analysis by discussing justification for statistical combination of the data (i.e. similarities and differences among the studies to be included in the meta-analysis)

IV. Evidence-Based Answer to Clinical Question

(i) Generate a consensus statement across the studies analyzed that specifically addresses and answers the research question, while clearly discusses the applications, implications and limitations of the findings

(ii) While systematic reviews are the best measures currently available to evaluate critically and to summarize data and support the effectiveness and efficacy of therapies, the success of this research lies in the stringent adherence to its protocols

Appendix 3: The Markov Process

Clinical decision-making problems often involve multiple transitions between health states. The probabilities of state transitions, or related utility values, require complex computations over time. Neither decision trees nor traditional influence diagrams offer as practical a solution as state of transition models (i.e. Markov models). This is so because Markov models are designed to represent cyclical, recursive events, whether short-term processes, and therefore are best used to model prognostic clinical cases, such as a surgical procedure and associated follow-up, or long-term management of a chronic disease, such as Alzheimer's disease, reliably and accurately. Markov models can be used to calculate a wide variety of outcomes, including average life expectancy, expected utility, long-term costs of care, survival rate and the number of recurrences.

Discrete Markov models enumerate a finite set of mutually exclusive possible states so that, in any given time interval (called a cycle or stage), an individual member of the Markov cohort can be in only one of the states. In order to determine a value for the entire process (e.g. a net cost or life expectancy), a value (an incremental cost or utility) is assigned to each interval spent in a particular state. The assignment of value in a Markov model is called a reward, regardless of whether it refers to a cost, utility or other attribute. A state reward refers to a value that is assigned to the members of the cohort in a particular state during a given stage. The actual values used for state rewards depend on the attribute being calculated in the model (e.g. cost, utility or life expectancy). A simple set of initial probabilities is used to specify the distribution of model subjects among the possible state rewards at the start of the process. The resulting matrix of transition probabilities is used to specify the transitions that are possible for the members of each Markov reward state at the end of each successive stage.

Two methods are commonly used to calculate the value of a discrete Markov model: (i) cohort (expected value) calculations and (ii) Monte Carlo trials. In a cohort analysis, which corresponds more realistically to a clinical situation, the expected values of the process are computed by multiplying the percentage of the cohort in a reward state by the incremental value (i.e. cost or utility) assigned to that state. The outcomes are added across all state rewards and all stages. In the more theoretical Monte Carlo simulation trial, the incremental values of the series of reward states traversed by the individual are summed.

The Markov model is most often represented in a graphical form known as a cycle tree. Since it is based on a node and branch framework, it is easily integrated into standard decision tree structures and can be appended to paths in a Markov decision tree. The root node of the Markov cycle tree is called a Markov node. Each of the possible health states is listed on the branches emanating from the Markov node, with one branch for each state. Possible state transitions are graphically displayed on branches to the right. A state from which transitions are not possible, such as the Dead state, is called an absorbing state. No state rewards are given for being in the Dead state, and zero values are assigned to the state rewards of all absorbing states. In this fashion, the Markov process integrates a termination condition, or stopping rule, specified at the Markov node to determine whether a cohort analysis is complete. This rule is the termination condition at the beginning of each stage. When the termination condition is verified, then the Markov process ends and the net reward(s) are

reported. The termination condition can include multiple conditions, which may be cumulative or alternative.

The Markov model generates an expected value analysis that is performed at or to the left of each Markov node in cohort analysis. The expected value analysis can generate additional information about the Markov cohort calculations. For example, in a model designed to measure the time spent in the diseased state diagnosed as dementia of the Alzheimer's type, an expected value will be generated to average life expectancy for a patient in the cohort. Additional calculated values will include the amount of time spent, on average, in each of the specified states of Alzheimer's dementia. The percentage of the cohort in each state will be computed at the end of the process. When the termination condition has been set to continue the process until most of the cohort is absorbed into the Dead state, the final probability of patients in the Dead state will approach 1.0. In brief, one of the strongest assets of the Markov model is its capacity to yield both an extensive numerical description of the process under study as well as a detailed graphical representation.