International Journal of Technology Assessment in Health Care, 16:1 (2000), 13–21.

EVALUATION OF RANDOMIZED CONTROLLED TRIALS ON COMPLEMENTARY AND ALTERNATIVE MEDICINE

Bernard S. Bloom
University of Pennsylvania

Aurélia Retbi
Université de Paris

Sandrine Dahan
Université Renée Descartes

Egon Jonsson
Karolinska Institute and The Swedish Council on Technology Assessment in Health Care (SBU)

Abstract

Objectives: Use of complementary and alternative medicine (CAM) is growing in all Western countries. The goal of this study was to evaluate quality of randomized controlled trials (RCTs) of CAM interventions for specific diagnoses to inform clinical decision making.
Methods: MEDLINE and related databases were searched for CAM RCTs. Visual review was done of bibliographies, meta-analyses, and CAM journals. Inclusion criteria for review and scoring were blinded RCT, specified diagnosis and intervention, complete study published between January 1, 1966 and July 31, 1998 in an English-language, peer-reviewed journal. Two reviewers independently scored each study.
Results: More than 5,000 trials were found, but only 258 met all study inclusion criteria. The main cause for rejection (> 90%) was that the study was not an RCT or had no blinding. Mean score across 95 diagnosis/intervention categories was 44.7 (S.D. ± 14.3) on a 100-point scale. Ordinary least-squares regression found date of publication, biostatistician as author or consultant, published in one of five widely read English-language medical journals and diagnosis/intervention category of hypertension/relaxation as significant predictors of higher scores.
Conclusions: The overall quality of evidence for CAM RCTs is poor but improving slowly over time, about the same as that of biomedicine. Thus, most services are provided without good evidence of benefit.

Keywords: Complementary medicine, Alternative medicine, Randomized controlled trials, Quality evaluation

The authors are indebted for assistance to Clifford S. Bayer, B.S., Douglas Bruno, B.A., Alan Kessler, B.S., Elizabeth R. Mackenzie, PhD, and Daniel Y. Maman, B.A., of the University of Pennsylvania. Funded in part by the Swedish Council on Technology Assessment in Health Care, Stockholm, Sweden.

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

There is increasing utilization of complementary and alternative medicine (CAM), alone or alongside biomedicine, in the United States (9), western Europe (12;21;26), and Australia (28). CAM is used in high-risk disease such as cancer (6), HIV/AIDS (2;36), and dementia (15), and low-risk diseases like those of the gastrointestinal tract (38). Services range from those requiring formal education, such as Ayurvedic, chiropractic, and acupuncture, to folk methods such as prayer. Changed perceptions in the United States about CAM are expressed in establishment in 1991of the Office of Alternative Medicine (OAM) in the National Institutes of Health, insurance company payment, and formal CAM curricula in about 50 medical schools in 1999 (none 20 years ago).

The goal of this study was to evaluate the clinical evidence from randomized controlled trials (RCTs) on CAM efficacy. Evaluating published studies informs clinical and economic decisions on appropriate use, resource allocation, and expected usefulness of interventions for patients (4).

Any review of quality of methodology and results requires comparable standards that CAM and biomedicine should meet (22;40). But the thousands of CAM studies published during the past three decades defy systematic review in an efficient manner. Thus, this study reviewed only published RCTs to establish a baseline confidence level for efficacy of defined CAM treatments for specified diagnoses.

The RCT is accepted as the best available method (by frequentist and Bayesian perspectives) to prove cause and effect while minimizing bias in observed (patient) and observer (practitioner, evaluator). But not everyone agrees the RCT is the best method to evaluate CAM (3;18;19;23;33). Similar arguments were made by biomedical practitioners after publication of the first modern RCTs in the 1940s, i.e., uniqueness of patient symptoms and diagnosis, inability of RCTs to test complexities of the healing process, and the need for individualized patient care. A. L. Cochrane did as much as anyone to support the value of RCTs (8).

Increased desire by patients for CAM (17), and payment in the United States (10;30), are complemented by physicians in many countries requesting more research into CAM efficacy. In most countries, e.g., China, France, Great Britain, Germany, India, and Sweden, coexistence of CAM and biomedicine is the norm. For example, use of St. John's wort to treat clinical depression is relatively widely accepted in western Europe, whereas it has only recently begun to receive serious attention in the United States.

METHODS

The definition of CAM used for this study was that of the Office of Alternative Medicine (OAM): ". . . those treatments and health care practices not taught widely in medical schools, not generally used in hospitals, and not usually reimbursed by medical insurance companies" (31).

Criteria for Study Collection

The published literature was obtained by 79 computer search terms from OAM, the Research Council for Complementary Medicine in London, and supplemented with terms found during computer database searching. Each term was of a specific intervention for a defined diagnosis. The review of MEDLINE and related systems used each diagnosis and/or intervention term matched with "randomized controlled trial" and its variations.

Five criteria were required for inclusion:

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

1. Prospective controlled trial with description of randomization;
2. Blinding of observers, observed, or both;
3. Specified diagnosis, treatment, and outcome;
4. Complete study, no abstracts, published in an English language, peer-reviewed journal; and
5. Published between January 1, 1966 and July 31, 1998.

Over 5,000 citations were reviewed. About 1,300 articles appeared to meet study criteria and were examined intensively. Over 75% of these 1,300 articles was rejected, leaving 258 RCTs reviewed and scored that serve as the basis of this study. Main reasons (> 95%) for rejection during screening processes were misclassification by MEDLINE and at least one inclusion criterion was not met; usually the study was not an RCT or had no blinding. A 10% random sample of rejected articles was rereviewed (by BSB) and all were correctly discarded (i.e., no false negatives). Only 25% of studies scored were found by all computerized searches combined.

Additional articles were found through the Cochrane Database of Systematic Reviews, bibliographies of publications, literature reviews and meta-analyses, ISI Citations Indexes, and visual search of selected CAM journals. The Centralized Information Service for Complementary Medicine (CISCOM) and EMBASE systems were not searched. There are few differences of results between MEDLINE and CISCOM searching systems (22).

Quality Review Process

A recent analysis of RCT systematic research scoring systems found all acceptable (29). We chose that of Chalmers et al. (7), which evaluates four areas of RCT quality:

1. Trial description;
2. Method—inclusion and exclusion criteria, withdrawals, interventions, blinding, sample size calculation, patient compliance, and side effects;
3. Statistical analysis—appropriateness, significance tests, inference, confidence limits, and regression analysis; and
4. Presentation of results.

Each study was reviewed and scored independently by two reviewers and given a numerical score. Reviewers were not blinded. An additional 4% of studies (false positives) was rejected, nearly always because they were not an RCT or had no blinding. Reviewers agreed on 94% of subitems scored. Scoring disagreements, mainly of interpretation of quality of study method or completeness of analysis, were resolved by discussion on every point. A 10% random sample was re-reviewed in blinded fashion by BSB, who found differences of no more than ± 2 points from the first review.

Data Analysis

Mean scores, percent distribution, and standard deviations were calculated. Diagnosis/intervention categories of $n > 10$ studies, or with $n = 5–10$ studies plus a coefficient of variation < 0.2, were also analyzed separately. Variables were tested by Student's t test for pair-wise comparisons and ANOVA.

An ordinary least-squares regression model estimated the effects on study score of diagnosis/intervention category, statistician as author or consultant, location of study (university, government institution, other), published in any of five widely read English language medical journals (*Annals of Internal Medicine, British Medical*

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

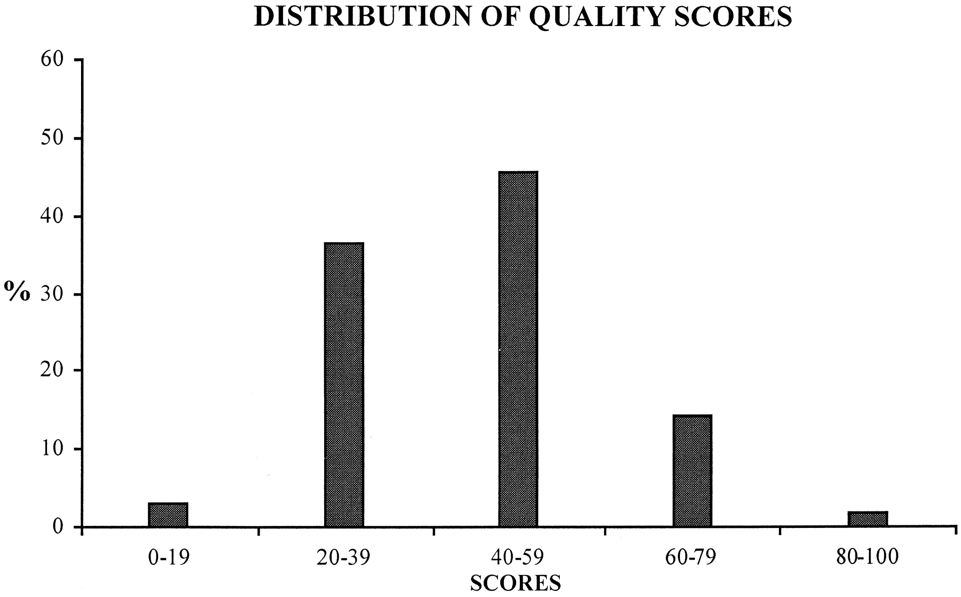

Figure 1. Distribution of quality scores.

Journal, JAMA, Lancet, or *New England Journal of Medicine*), year of publication, statistical significance of endpoint(s), and source of funding (government, private, other, none).

RESULTS

Quality Scores

There were 95 separate diagnosis/intervention categories from the 258 studies scored, 59 (62.8%) of which had only one study. Mean score on a 100-point scale across all CAM trials was 44.7 (S.D. ± 14.29), range 10–93 (Figure 1). This review's quality scores were similar to those of an evaluation of biomedicine RCTs (45 points) (11).

Only 1.2% of studies had scores of 80 points or greater. Examples included zinc gluconate for the common cold, transcutaneous electrical nerve stimulation for chronic low back pain, and homeopathy for upper respiratory infection and seasonal allergies. At the other extreme, 3.0% of studies had a score less than 20 points. We chose 60 points as the minimum acceptable quality score, the same minimum passing grade of many schools; 15.3% of evaluated studies had scores of 60 points or greater. The most common deficiencies were no sample size calculation ($> 90\%$), insufficient blinding of patient, provider, evaluator or statistician ($> 90\%$), no patient reject or withdrawal information ($> 90\%$), no testing of blinding, randomization, or compliance ($> 90\%$), and inadequate statistical analysis (75%).

Analysis of independent variables found similar results by Student's t test and ANOVA. Independent variables significantly associated with higher quality scores were: a) having a biostatistician as either an author or consultant ($t = -4.55$, $p = .00001$; ANOVA F = 10.52, $p = .0001$); b) any external funding (public or private) ($t = -3.23$, $p = .0001$; ANOVA F = 5.53, $p = .005$); c) published in any of five English language medical journals ($t = -4.19$, $p = .00001$; ANOVA F = 17.58,

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

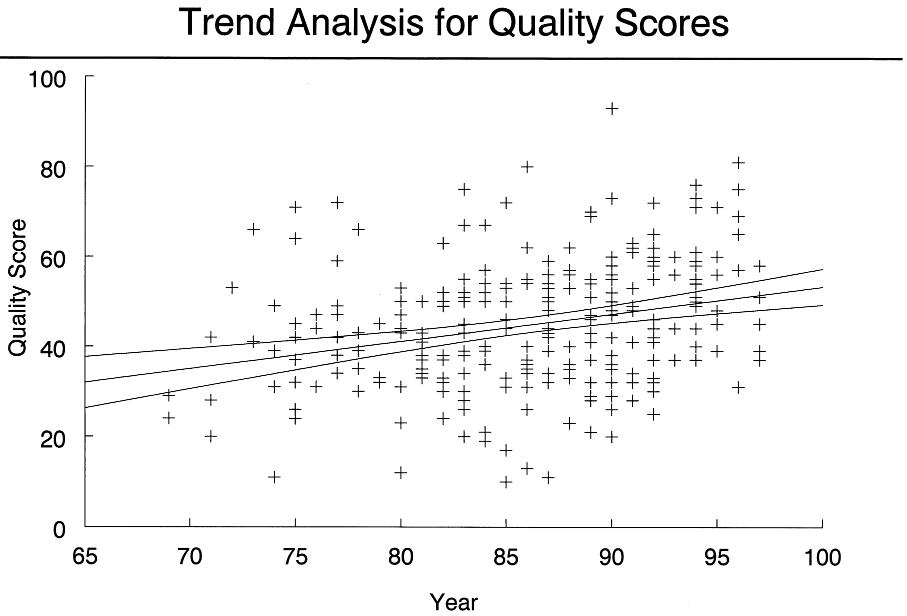

Figure 2. Trend analysis of quality scores, 1966–97.

p = .0001); and d) having a positive or significantly positive result for the study intervention ($t = 2.72, p = .007$; ANOVA F = 3.68, p = .02). Study location (private, university or government hospital, or ambulatory facility) was not significant (t = −1.08 to 1.54, p = .07 to .28; for all pairwise comparisons ANOVA F = 1.82, p = .16).

Trend analysis found significantly increasing quality scores over time (Figure 2). Studies published more recently had higher mean quality scores than those published earlier, but the distribution of deficiencies did not change over time: y-intercept = 34.4; slope = .061, 95% C.I. (around the annual mean score of 44.7) = 35–87; S.E. = .13; and t = 4.6, p = .00001.

Six diagnosis/intervention categories allowed specific analysis, two with n > 10 studies and four with n = 5–10 studies plus a coefficient of variation less than 0.2 (Table 1). Mean and distribution of quality scores for this group were similar to those of all other studies.

Modeling Results

Ordinary least-squares regression found four variables as significant predictors of quality score (R^2 = 19.6%): a) date of publication; b) biostatistician as author or

Table 1. Selected Quality Scores: n > 10 Studies, or n = 5–10 Studies Plus Coefficient of Variation < 0.2

Intervention	Diagnosis	n	Mean	± S.D.
Spinal manipulation	Back/neck pain	45	39.2	12.3
Relaxation	Hypertension	22	49.0	12.0
Electrosleep	Depression	8	45.5	5.8
Yohimbe	Male sexual dysfunction	7	38.9	5.6
Garlic	Elevated lipids	6	45.9	8.7
TENS	Osteoarthritis pain	5	37.2	7.0
All studies			44.7	14.3

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

Table 2. Results of Ordinary Least-squares Regression Model[a]

Variable	Coefficient	95% C.I.	*p* Value
Publication date	0.63	0.35–0.92	.0001
Biostatistician	6.71	2.29–11.14	.033
English language journal	6.68	1.55–11.81	.012
Hypertension/relaxation	7.75	1.44–13.68	.016

[a] $R^2 = 19.65\%$.

consultant; c) published in one of the five selected English language journals; and d) diagnosis/intervention category of hypertension/relaxation (Table 2). However, more than 80% of the variation in scores was not explained by this regression model.

DISCUSSION

Most CAM modalities pre-date contemporary biomedicine, often by thousands of years. Medicinal herbs and prayer have been used by virtually every human culture, while formal healing systems like Ayurvedic and traditional Chinese medicine are at least several thousand years old. Thus, a large body of anecdotal evidence and cultural memory is available in evaluating CAM (16). But, as this study shows, nearly all CAM interventions examined are of unknown benefit.

Acceptance of contemporary methodological standards is rooted in underlying philosophies of evidence and usefulness. Biomedicine has accepted RCTs as the highest current methodologic standard, which is also increasingly being applied in CAM. For example, the Cochrane Database of Systematic Reviews has published reviews of RCTs of specific CAM modalities. Meta-analyses are also increasingly used in CAM, like those of homeopathy (24;27) and spinal manipulation for low back pain (32;35).

Acceptance of CAM by patients, providers, and payers is likely to increase use further (9;10) without increasing knowledge of efficacy. For example, widespread over-the-counter availability in many countries of nutritional supplements makes evaluation exceedingly difficult. While biomedical journals are increasingly publishing CAM studies, for example, of laughter (42) and touch (34), CAM journals face restricted distribution and unavailability in most databases used by biomedicine physicians.

RCT quality is equally poor in biomedicine and CAM, and thus there is little *a priori* confidence of population benefit for most interventions. But CAM study quality increased significantly over time, as undoubtedly has that of biomedicine. However, more biomedicine interventions have been subjected to RCTs than have CAM treatments.

There are ways to overcome deficiencies. First is to undertake short- and long-term randomized controlled cost-effectiveness trials evaluating CAM and conventional medical care for specific diagnoses (39). Second is to undertake meta-analyses of existing RCTs. But meta analyses do not replace RCTs, as they may show similar (5) or different results (24;25;39;41) when compared to large or small RCTs (20). A third alternative, which is common policy in the United States, is to do nothing and let the market decide what is provided and who pays.

Reliance on the economic market, though, is also likely to have little positive effect on CAM use. We know the economic market for health and medical care services works poorly — it is not the same as that for financial services and stereo

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

equipment where most people pay for goods or services directly and can determine value (benefit-cost tradeoffs among alternatives). But in health care, a different order rules — information asymmetry under conditions of uncertainty and first-dollar payment by insurers mean many are happy to have others pay for their desires and physician decisions, substituting belief of possible benefit for scientifically derived knowledge of benefit.

Whether a society wants and is willing to pay for parallel systems of health and medical care (e.g., medical pluralism) (14), with many based on traditions and one on contemporary scientific standards, warrants further public discussion. Interventions based on biomedical or CAM philosophies of health and illness are difficult to justify, given that a very large percent of both have never been tested adequately and are of unknown value.

None of this is to suggest that tradition should be ignored. There are examples of ancient treatments that are still used, such as successful trephination of the skull in pharaonic Egypt, leeches to aid wound healing, and bloodletting for "dropsy" (congestive heart failure). The current "gold standard" research method, the RCT, is itself not new. For example, the tenth century Persian polymath ibn Sina (Avicenna) recommended randomizing patients to various medications to test outcomes (14), as did van Helmont in the early 17th century to evaluate bloodletting (37). Such examples, though, must be weighed against the larger array of failures, many of which were not only subsequently proven useless, but were often dangerous, like frequent bleeding and purging, and in this century, gold for treating tuberculosis (1).

Uncritical acceptance of any medical philosophy or untested treatment may increase personal freedom and individual choice, but it also means patients and practitioners are confronted by an array of uncertain medical care choices and unpredictable health and economic consequences. If individuals are willing to pay directly, they should have the opportunity to do so. But all high-income countries pay for all or most medical care through taxes and social insurance, and there should be reasonable certainty that they work, whether conventional or CAM interventions.

Study Limitations

The first limitation of this study is that the evaluation was only of the published article, not the actual study. A low evaluation score may be due to poor method or analysis, inadequate reporting by the authors, or space limits by the journal.

Second, there is a substantial literature in languages other than English, such as Chinese, French, German, Hindi, Italian and Russian. Including these studies would have increased the number of RCTs reviewed and may have influenced the results.

Next, most computerized databases do not adequately collect and categorize CAM publications. For example, only 3 of 15 chiropractic journals are available on MEDLINE. However, visual review of CAM journals produced few additional trials.

Fourth, any qualitative review may have its own biases. For example, interrater reliability was not tested because there was 94% agreement of item scores.

Last, study quality is independent of whether an intervention works. Unproven is not synonymous with not efficacious.

CONCLUSION

The primary influence of research publications is to inform decisions by patients, practitioners, payers and policy makers. Inadequate studies have no value, since

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

they may lead to overuse of ineffective and underuse of effective technology by clinicians (40).

The breadth of RCTs reviewed across many interventions and diagnoses means that one can place a high level of confidence in the primary finding of this study—that the quality of evaluated studies was low, with uncertain benefits of nearly all CAM interventions. There is equal concern of quality in biomedicine. While far more biomedicine interventions have been rigorously studied than those of CAM, mean quality scores were essentially the same. Thus, inadequate evidence of benefit afflicts both forms of medicine.

REFERENCES

1. Amberson, J. B., McMahon, B. T., & Pinner, M. A clinical trial of sanocrysin in pulmonary tuberculosis. *American Review of Tuberculosis*, 1931, 24, 401–35.
2. Anderson, W., O'Connor, B. B., MacGregor, R. R., & Schwartz, J. S. Patient use and assessment of conventional and alternative therapies for HIV infection and AIDS. *AIDS*, 1993, 7, 561–66.
3. Anthony, H. M. Some methodological problems in the assessment of complementary therapy. *Statistics in Medicine*, 1987, 6, 761–71.
4. Bloom, B. S., & Fendrick, A. M. Timing and timeliness in medical care evaluation. *PharmacoEconomics*, 1996, 9, 183–87.
5. Cappelleri, J. C., Ioannidis, J. P. A., Schmid, C. H., et al. Larger trials vs meta-analysis of smaller trials: How do their results compare? *JAMA*, 1996, 276, 1332–38.
6. Cassileth, B. R., & Chapman, C. C. Alternative and complementary cancer treatments. *Cancer,* 1996, 77, 1026–34.
7. Chalmers, T. C., Smith, Jr., H., Blackburn, B., et al. A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, 1981, 2, 31–49.
8. Cochrane, A. L. *Effectiveness and efficiency: Random reflections on health services.* London: The Nuffield Provincial Hospitals Trust, 1972.
9. Eisenberg, D. M., Davis, R. B., Ettner, S. L., et al. Trends in alternative medicine use in the United States, 1990–1997. *JAMA*, 1998, 280, 1569–75.
10. Elder, N. C., Gillcrist, A., & Minz, R. Use of alternative health care by family practice patients. *Archives of Internal Medicine*, 1997, 6, 181–84.
11. Emerson, J. D., Burdick, E., Hoaglin, D. C., Mosteller, F., & Chalmers, T. C. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials*, 1990, 11, 339–52.
12. Fisher, P., & Ward, A. Complementary medicine in Europe. *British Medical Journal*, 1994, 309, 107–11.
13. Good, C. M. Medical pluralism. In R. Ornstein & C. Swencionis (eds.), *The healing brain: A scientific reader.* New York: The Guilford Press, 1990.
14. Green, F. H. K. The clinical evaluation of remedies. *Lancet*, 1954, 2, 1084–91.
15. Hogan, D. B., & Ebly, E. M. Complementary medicine use in a dementia clinic population. *Alzheimer Disease and Associated Disorders*, 1996, 10, 63–67.
16. Hufford, D. J. Contemporary folk medicine. In N. Gevitz (ed.), *Other healers*. Baltimore: The Johns Hopkins University Press, 1988.
17. Hufford, D. J. Cultural and social perspectives on alternative medicine: Background and assumptions. *Alternative Therapies*, 1995, 1, 53–61.
18. Hufford, D. J. Culturally grounded review of research assumptions. *Alternative Therapies*, 1996, 2, 47–53.
19. Hufford, D. J. Integrating complementary and alternative medicine into conventional medical practice. *Alternative Therapies*, 1997, 3, 81–83.
20. Ioannidis, J. P. A., Cappelleri, J. C., & Lau, J. Issues in comparisons between meta-analyses and large trials. *JAMA*, 1998, 279, 1089–93.

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press

21. Johannessen, H., Launso, L., Olesen, S. G., & Staugard, F. (eds). *Studies in alternative therapies, I: Contributions from the Nordic countries.* Gylling, Denmark: INRAT/Odense University Press, 1994.
22. Jonas, W. B. Researching alternative medicine. *Nature Medicine*, 1997, 3, 824–27.
23. Kiene, H. & von Schon-Angerer, T. Single-case causality assessment as a basis for clinical judgement. *Alternative Therapies*, 1998, 4, 41–47.
24. Kleijnen, J., Knipschild, P., & ter Riet G. Clinical trials of homeopathy. *British Medical Journal*, 1991, 302, 316–23.
25. LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. Discrepancies between meta-analyses and subsequent large randomized controlled trials. *New England Journal of Medicine*, 1997, 337, 536–42.
26. Lewith, G., & Aldridge, D. *Complementary medicine and the European community.* Saffron Walden, Finland: C. W. Daniel, 1991.
27. Linde, K., Clausius, N., Ramirez, G., et al. Are the effects of homeopathy placebo effects? A meta-analysis of placebo-controlled trials. *Lancet*, 1997, 350, 834–43.
28. MacLennon, A. H., Wilson, D. H., & Taylor, A. W. Prevalence and cost of alternative medicine in Australia. *Lancet*, 1996, 347, 569–73.
29. Moher, D., Jadad, A. R., Nichol, G., et al. Assessing the quality of randomized control trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials,* 1995, 16, 62–73.
30. Moore, N. G. A review of reimbursement policies for alternative and complementary medicine. *Alternative Therapies*, 1997, 3, 26–29, 91–92.
31. National Institutes of Health, Office of Alternative Medicine. *Alternative medicine: Expanding medical horizons (A report to the National Institutes of Health on alternative medical systems and practices in the United States).* Bethesda: NIH, 1992.
32. Ottenbacher, K., & DiFabio, R. P. Efficacy of spinal manipulation/mobilization therapy: A meta analysis. *Spine*, 1985, 10, 833–37.
33. Patel, M. S. Problems in the evaluation of alternative medicine. *Social Science and Medicine*, 1987, 25, 669–78.
34. Rosa, L., Rosa, E., Sarner, L., & Barrett, S. A close look at therapeutic touch. *JAMA*, 1998, 279, 1005–10.
35. Shekele, P. G., Adams, A. H., Chassin, M. R., Hurwitz, E. L., & Brook, R. H. Spinal manipulation for low-back pain. *Annals of Internal Medicine*, 1992, 117, 590–98.
36. Singh, N., Squier, C., Sivek, C., et al. Determinants of nontraditional therapy use in patients with HIV infection. *Archives of Internal Medicine*, 1996, 156, 197–201.
37. Van Helmont, J. A. John Chandler, translater. *Oriatrike, or physick refined. The common errors therein refuted and the whole are reformed and refined.* London: Loyd, 1662.
38. Verhoef, M. J., Sutherland, L. R., & Brkich, L. Use of alternative medicine by patients attending a gastroenterology clinic. *Canadian Medical Association Journal*, 1990, 142, 121–25.
39. Vickers, A., Cassileth, B., Ernst E, et al. How should we research unconventional therapies? *International Journal of Technology Assessment in Health Care*, 1997, 13, 111–21.
40. Vickers, A. J. Can acupuncture have specific effects on health? A systematic review of acupuncture antiemesis trials. *Journal of the Royal Society of Medicine*, 1996, 89, 303–11.
41. Villar, J., Carroll, G., & Belizán, J. M. Predictive ability of meta-analyses of randomised controlled trials. *Lancet*, 1995, 345, 772–76.
42. Yoshino, S., Fuhimori, J., & Kohda, M. Effects of mirthful laughter on neuroendocrine and immune systems in patients with rheumatoid arthritis. *Journal of Rheumatology*, 1996, 23, 794–98.

https://doi.org/10.1017/S0266462300016123 Published online by Cambridge University Press