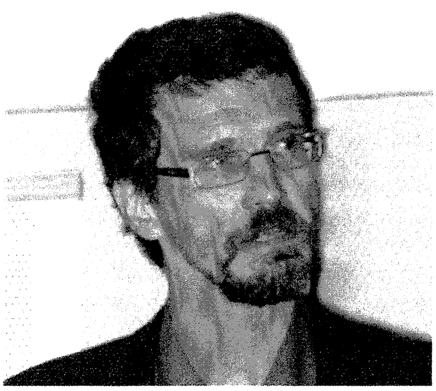

Harri Hemilä, MD, PhD

Limitations of Randomized Trials: Problems in the "Evidence" of EBM

Introduction

This commentary is motivated by the adherence of Evidence-based Medicine (EBM) to two principles: 1) the restriction of the evaluation of therapeutic effects to randomized trials (RCT) and 2) the proposal that ordinary clinicians should search for and read and interpret RCTs by themselves.

EBM was proposed in 1992 in a paper published in JAMA[1]. That paper stated that the primary methodological criterion for a treatment trial should be to ask: "Was the assignment of patients to treatments randomized?" Strict focus on randomized studies has also been promulgated in other texts on EBM, such as the EBM textbook by Sackett et al.[2], which stated that "If you find that the study was not randomized, we'd suggest that you stop reading it and go to the next article." Furthermore, Cochrane collaboration advertises that they "are world leaders in evidence-based health care"[3] and in the evaluation of treatment effects, the Cochrane collaboration is principally restricted to RCTs.

The second principle of EBM was formulated in the 1992 paper as follows: "The underlying belief is that physicians can gain the skills to make independent assessments of evidence" and the paper explicitly put "a low value on authority" ([1], p. 2421). "Asking a local expert" was described as a typical approach of old-fashined medicine, which is inconsistent with the EBM requirement that each physician should assess the evidence for himself or herself[1]. Finnish proponents of EBM formulated this principle by way of an example: "In the era of electronic information a doctor in Utsjoki in Lapland can read the same newly published paper online that an Oxford professor uses when teaching that same afternoon" ("Sähköisen tiedonvälityksen aikakaudella Utsjoen lääkäri voi aamulla lukea tuoreeltaan verkkojulkaisun, johon Oxfordin professori iltapäivällä opetuksessaan viittaa")[4].

Thus, in plain language, the perception of EBM is that upon waking the GP should first go to the internet to find out what new RCTs have been published to update his or her methods of treatment. Furthermore, when a patient complains of "back pain", the GP should start the encounter by first searching and reading what is "the latest RCT evidence" for back pain treatment.

Previously, a few Finnish papers discussed various problems with the EBM approach[5-12]. Citations of the English literature that criticizes EBM are not listed here, but the Finnish papers guide to some relevant papers. This chapter summarizes some problems encountered with the EBM proposal by which ordinary clinicians should routinely search and read RCT reports.

Problems with the proposal that ordinary physicians should routinely search and read RCTs

1. The numbers of RCTs is huge

The number of published RCTs is so large that no person can read RCT literature of popular topics. For example, in August 2012, Medline data base contained 10 804 RCT reports for "NSAID", 1761 RCT reports for "back pain" and 120 for "back pain & NSAID" [13]. For "hypertension", Medline contained 10 215 RCT reports.

The problem with the large volume of data is not restricted to the cumulative information, but new RCTs are published at a high rate. Bastian et al. calculated that a mean of 75 new RCTs and 11 new meta-analyses are published every day[14]. If one would like to peruse such a large number of new reports even at the level of abstracts, it would take all day and a clinician would have no time for treating patients.

Because of the large number of RCTs, the 1992 EBM paper stated that "Meta-analysis is gaining increasing acceptance as a method of summarizing the results of a number of RCTs" [1]. However, the number of meta-analyses is also so large that a clinician cannot read them either. For example, there were 740 meta-analyses for "NSAID", 176 for "back pain", and even the more selective combination of "back pain & NSAID" identified 8 meta-analyses in the Medline data base[13]. For "hypertension", Medline identified 795 meta-analyses.

Thus, it is impossible for the GP in Lapland to comply with the EBM suggestion that an ordinary clinician should search and read the RCT literature for back pain when a patient complains of back pain. Furthermore, major journals are not free and only a limited number of journals are available for an ordinary GP. Journals are largely owned by commercial companies and the establishment of the internet has not changed the ownership characteristics.

2. The results of different RCTs are often inconsistent

The RCT is usually considered as the "gold standard" for the evaluation of treatment effects, which is the reason for EBM to restrict the appraisal of treatment studies to RCTs. However, the results of RCTs on the same topic are often contradictory. Furukawa et al. searched for pairs of large trials on the same treatment and found 289 pairs. As many as 27 per cent of those pairs reported significantly different results[15]. Thus, the "gold standard" status of RCT is limited by such contradictions.

There are a number of evident explanations for the inconsistent RCT results, even though they nominally test "the same treatment". For example, the study participants differ and the definition of disease and its severity are not identical from one study to another. Even though the treatment is classified as the same, there can be differences in the dose, duration and other methodological aspects, and there can be differences in the co-treatments. Finally, outcomes are often pragmatic, differing between studies and therefore RCTs on the same disease can measure somewhat different effects. Minor to moderate differences in the above issues can add up to substantial variation in the observed effects even though two RCTs are thought to evaluate "the same treatment".

Such differences between RCTs make it difficult to generalize the findings of a single RCT. For example, if an RCT was carried out in India or Hungary, is it relevant or irrelevant for the GP working in Finnish Lapland? Of course, there is no universal answer to such a question, since generalization depends on the disease and the treatment, and on the cultural context and the health care system, etc.

Meta-analysis has been suggested as a solution to the large numbers of RCTs and their diverse findings[1]. However, the findings of a meta-analysis depend on the selection criteria used for the RCTs that are analyzed. Setting the selection criteria is a subjective issue. Variation in the selection criteria can lead to substantially different conclusions, even though two discrete but similar meta-analyses examine "the same treatment". For example, Prins et al.[16] analyzed 4 different meta-analyses on the role of administration frequency of aminoglycosides and found that all 4 meta-analyses drew different conclusions. This was explained by "... variation in the study selection criteria applied. Therefore, the number of studies included varied from 13 to 24..." The conclusion of these authors was cynical: "The physician can only follow the conclusion of the meta-analysis most closely in accordance with his or her own beliefs" (sic)[16].

As a further source of confusion, meta-analyses of small trials have often produced results inconsistent with those of large single trials[17].

Finally, there are many examples of meta-analyses in which the figures extracted from the RCTs are incorrect and/or the calculation methods are unsound[18,19]. A recent Cochrane meta-analysis on zinc treatment for the common cold[20] is a good example of serious problems that sometimes occur in meta-analyses. In a four-page critical feedback to the review, Hemilä pointed out that the authors had excluded some trials by using unsound arguments and there were errors in the description of the trials and in the data extraction. Moreover, trials that were very different were nevertheless pooled (the apples and oranges problem), the scale used for combining results was unsound, etc.[21]. So far no correction or reply to the feedback has been published. Given that Cochrane collaboration advertises itself as the "world leaders of EBM", such a poor quality meta-analysis casts doubt on the validity of some of their meta-analyses. Some Cochrane reviews are good, but others are not. Unfortunately, the reader must be an expert to evaluate whether a meta-analysis is sound or not.

Thus, even though the GP in Lapland would have spent much time for searching and had found RCT reports or a meta-analyses on back pain, it is not obvious that they are relevant for the patient.

3. RCT gives an average effect that describes the population as a whole, but clinicians treat individual patients

An RCT determines whether there is an average difference between the treated and the control patients. However, patients can differ markedly from each other. Some of them get greater than average improvement whereas some may get no benefit at all (Fig. 1). Clinicians treat individual patients, and therefore "the average effect" found in an RCT has often little value for the clinician. The relevant question for a clinician is "Does the treatment help Mr. Smith or Mr. Jones?" instead of pursuing the average effect in an RCT which describes a whole population.

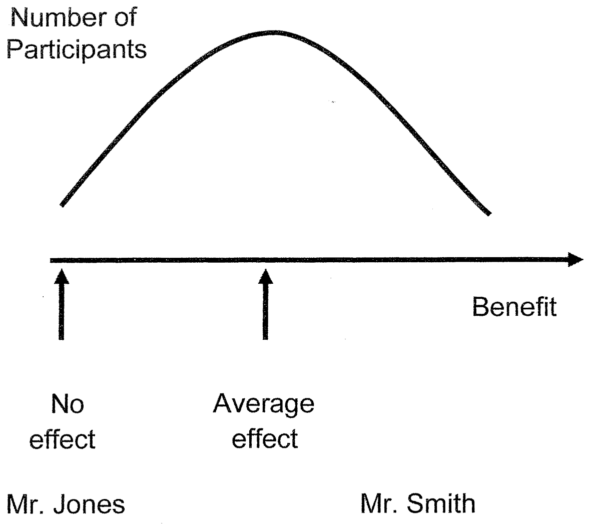

Fig 1.

In the RCT field, variation of the treatment effect over study population is called "subgroup differences" or "within-study heterogeneity". Such variation can be studied in big RCTs, whereas small RCTs do not have sufficient statistical power to analyze subgroup differences. Subgroup analysis can get us closer to the individual we are treating.

As an example of considerable subgroup differences for a treatment effect, Hemilä and Kaprio analyzed the effect of vitamin E supplementation on mortality in the large Finnish ATBC Study on 29 133 male smokers aged 50 to 69 years[22]. Vitamin E had no overall effect on mortality with a risk ratio (RR) = 1.02. However, previous work on ATBC Study infections found significant heterogeneity in the vitamin E effect between subgroups and this necessitated subgroup analysis on mortality. Table 1 shows significant variation in the vitamin E effect on mortality when the ATBC participants were analyzed by age and dietary vitamin C intake (cut at the median).

	Age of the participant		
	50-62 y	63-65 y	66-69 y
Dietary vitamin C	Effect of vitamin E on the mortality rate		
Low	0%	-5%	+7%
High	+19%	-11%	-41%

Table 1: Effect of vitamin E administration on mortality in ATBC Study subgroups[22]

The average effect of vitamin E supplementation over all ATBC participants was +2% (RR = 1.02). Test of heterogeneity over the six subgroups: P = 0.0005. In the three subgroups with low vitamin C intake, there is no difference between vitamin E and placebo.

Thus, RR = 1.02 is a statistically valid average effect for the whole ATBC Study population. However, on the basis of the data shown in Table 1, close to half of the ATBC participants - the old and the young with high vitamin C intake - were inconsistent with the average of RR = 1.02. Thus, the average effect of a treatment can be misleading for many individuals of the population.

The clinician is interested in the individual patient, whereas most RCTs give only the average effect. RCTs are used to study many different research questions. In some cases the average at the population level is relevant and sufficient information for guiding further policy. Nevertheless, the average effects can have minimal importance for the clinician. To illustrate the diversity of RCTs, let us consider two extreme examples:

1) Long term risk reduction trials: death, stroke, tuberculosis, etc.
In these kinds of RCTs we cannot ask the patient himself or herself about the benefit. Instead, we are restricted to the "average effect" which gives the justification for public health level policies. For example, policies for using aspirin for the secondary prevention of cardiovascular diseases or for giving various vaccinations are based of RCTs. However, the RCT reports on these issues are of little relevance to the ordinary clinician. Instead, such RCTs should be interpreted by specialists, so that the specialists can instruct wider circles of ordinary physicians by the way of recommendations or enforce them through legislation. However, this is inconsistent with the EBM principle of putting "a low value on authority" ([1], p. 2421).

2) Short term subjective benefit RCTs: Does an NSAID help against back pain?
In these kinds of RCTs, we can ask the patient about the benefit of the treatment. When treating the individual, the size of the "average effect" is usually irrelevant after the treatment has been shown to be more effective than the placebo.
Let us consider patients Mr. Smith and Mr. Jones (Fig. 1) from the point of view of this latter type of RCT. Let us assume that both patients suffer from back pain and they cannot sleep properly because of it. We prescribe both of them the same NSAID. On his second visit, Mr. Smith recounts that the small dose completely relieved his back pain so that he can now sleep well. As clinicians, we interpret that the benefit is caused by the pharmacological effects of the NSAID, although we can keep in our minds that some part of the benefit might also be caused by the placebo effect.

On his second visit, Mr. Jones states that even at the highest allowed dosage the NSAID had no effect on his back pain. We do not start arguing with Mr. Jones that good quality RCTs have proven that the NSAID is effective and therefore he must be wrong. Instead, we trust the reporting of Mr. Jones and as clinicians we interpret that, as an individual case, this patient is far from the average effect of the drug (see Fig. 1).

Thus, as clinicians, we are continuously doing experiments at the individual level when treating these kinds of patients with back pain. As an analogy, antihypertensive drugs are tested individually for finding out which one of them works and what is the efficacious dose, instead of selecting a fixed drug and dose for all hypertensive patients on the basis of average effects found in RCTs. Many encounters with patients fall to this category; i.e., treat and look for the progress of the patient over time. Therefore, the average effect found in short term subjective benefit RCTs is mostly useless for an ordinary clinician.

As shown above, the literature on "back pain & NSAID" is extensive with 120 RCTs and 8 meta-analyses. Might reading these papers change the treatment of Mr. Smith and Mr. Jones, as the EBM proponents suggest? Given the individual variability on the effects of NSAIDs, it seems highly unlikely that reading the RCT literature would substantially change the treatments of Mr. Smith and Mr. Jones. The time of clinicians is limited and reading the 128 papers would leave numerous patients without treatment; i.e., all of them who could be treated during the time used for reading the 128 papers.

In the case of NSAIDs for treating back pain, the RCT data are relevant for the regulatory authorities and specialists of back pain, but ordinary clinicians get little or no advantage from reading the RCT reports when treating individual patients.

4. The EBM dogma: Observational studies are untrustworthy

As stated in the Introduction, the EBM textbook written by Sackett et al. suggested that "If you find that the study was not randomized, we'd suggest that you stop reading it and go to the next article"[2]. In effect, such policy would ignore all the studies on smoking and alcohol. Those authors' suggestion is based on the EBM belief that observational studies (case-control and cohort studies) are inherently untrustworthy.

There are a few systematic comparisons of RCTs vs. case-control and cohort studies on the same topics. Vandenbroucke[23] summarized the comparisons in a recent BMJ editorial as follows: "empirical proof that observational studies of treatment are widely off the mark has been surprisingly elusive. Four meta-analyses contrasting RCTs and observational studies of treatment found no large systematic differences." Vandenbroucke concluded that "the notion that RCTs are superior and observational studies untrustworthy ... rests on theory and singular events - discrepancies in the effects of vitamins." Thus, the substantial discrepancies seen in the cohort studies and the RCTs on some topics such as vitamin E should not be interpreted as an evidence against relying on case-control and cohort studies in general. On many topics, firm treatment conclusions can be drawn from observational studies, the harm done by smoking and drinking too much of alcohol being good examples. Thus, the EBM dogma that observational studies are inherently untrustworthy is false.

Placebos are commonly used in RCTs because their use decreases the risk of bias in studies. However, the importance of placebo depends on the measured outcome. A comparison of pharmacological placebo group with a no-treatment group (mostly 3-arm trials with the active treatment as the 3rd arm) found no differences when the outcomes were dichotomic (yes or no)[24]. However, when the outcome was continuous, there were significant differences between the placebo and no-treatment groups, and the evidence was particularly strong in 60 studies on pain[24].

Relevant methods depend on the specific topic. For example, case-control studies are useful when studying harmful effects of drugs[25] and cohort studies are useful for studies on smoking and alcohol. Placebo is usually essential when an RCT investigates pain. In contrast, the use of a placebo is not essential for RCTs with objective and dichotomic outcomes. Finally, in some cases even case reports are informative[26].

As a method, RCT is most relevant when a disease is common, treatments are expensive or long, effects are small, outcomes are subjective (pain) and there are strong financial interests. Back pain and hypertension are examples of topics for which RCTs are relevant. However, this does not mean that an ordinary clinician should start reading the 10 000 RCTs on hypertension and the 2000 RCTs on back pain.

5. EBM encourages critical thinking

The original EBM paper (1992) stated that "The underlying belief is that physicians can gain the skills to make independent assessments of evidence"[1]. This view is unrealistic. It is easy to read an abstract of an RCT report, but to understand the limitations of a single RCT requires much time and competence. Even more time is needed to understand why several RCTs on the same topic disagree and to decide which one of them is most relevant.

EBM gives guidelines for reading papers, but usually they transform complex issues into simplistic black and white issues. A good example of this phenomenon is the original proposal to look at "Was the assignment of patients to treatments randomized?"[1].

Proponents of EBM, including the Cochrane collaboration, have suggested Yes-No lists for various study quality items including: allocation concealment, randomization, blinding, and intention-to-treat (ITT), etc. However, going through an RCT report with such a list is not critical thinking, but it is "ticking the boxes". Most of these issues are complex. For example, alternative allocation is not randomization, but if patients are divided into study groups by their odd or even dates of birth, there is no basis to assume that the groups are systematically unbalanced. Thus, "not randomized" does not directly imply poor validity of methodology.

Furthermore, the importance of the particular quality item often depends on the observations. If the finding is positive, it is possible to explain the finding by the lack of placebo. However, if the finding is negative, usually it is not reasonable to explain the observation by the lack of placebo. Thus, "lack of placebo" has different importance depending on the actual observations. "Lack of placebo" does not always make studies untrustworthy, even if the study topic might be pain.

It seems probable that ordinary clinicians would eventually gain the skills to read original RCT reports critically, but they would need to use much time for courses on clinical epidemiology. Nevertheless, it seems questionable whether the time-benefit ratio for such an education is feasible. Being able to read original RCT reports does not mean that the clinician does actually have the time to read them or that the reading of RCTs would have any meaningful effect on the treatment of individual patients (see comments 1 to 4 above). Simplistic guidelines, such as look at whether a study used randomization or not, do not make ordinary physicians critical RCT report readers although EBM suggests otherwise[1,2].

6. Focus on RCTs can lead to bias in the evaluation of treatments

Pharmaceutical companies have the financial and organizational resources to carry out high-quality RCTs and thus they can hire top level statisticians for planning and analyzing RCTs. Drugs are well standardized products, which helps the generalization of findings. New drugs are patented so that an RCT is a good investment for the company assuming that the finding is positive.

In the case of non-pharmaceutical treatments, money is much more limited. Furthermore, non-pharmaceutical treatments are often not well standardized. For example, if one method of manual therapy is shown to be effective, it is not clear how far the finding can be extrapolated to other forms of manual therapy. There are many conceptional problems in the RCTs on psychotherapies in which the physician and the patient have very close interaction[8-10]. Similar problems may also be valid for RCTs on manual therapies. Thus, the EBM focus on RCT causes a strong bias towards drug treatments.

This comment does not oppose the method of RCT per se. As noted at the end of comment 4, there are many conditions when RCTs are particularly important. However, this does not mean that ordinary clinicians should routinely use their time for searching and reading RCT reports in their clinical work. Instead "asking a local expert" and putting a high "value on authority" - who have time and the competence to evaluate whether and how a new RCT contributes to the body of knowledge - seem much more practical approaches for the clinician. However, such approaches are explicitly discouraged by EBM[1,4].

Summary

Proponents of EBM suggest that ordinary physicians should search and read RCTs themselves[1,4]. However, this suggestion has numerous problems:
1. The number of RCTs is huge and no clinician has time to read them systematically, except when one limits the literature search to a very narrow medical question.

2. The findings of RCTs are sometimes inconsistent and it is not clear which one of the conflicting RCTs should be trusted. It is not always clear how far the findings can be generalized.
3. The average effect obtained in an RCT does not apply to all individuals and therefore the average is often useless for the clinician treating an individual patient.
4. RCT is not the only way to get valid information about treatment effects. Case-control and cohort studies can yield valid information and they should not be classified as inherently untrustworthy.
5. Ordinary clinicians do not have the competence to interpret RCT reports, which often are highly technical. Much formal education on clinical epidemiology is needed to make an ordinary clinician competent to critically read RCT reports, but that would mean treating fewer patients.
6. Strong focus on RCTs may cause strong bias in favour of drug therapy approaches.

References

1. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2420-5.

2. Sackett DL, Richardson WS, Rosenberg WS, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. NY: Churchill Livingstone 1997, p. 94.

3. Cochrane Collaboration. http://www.cochrane.org/

4. Mäkelä M, Kaila M. Tietoa järjestelmällisesti, hoitoa humaanisti. Duodecim 2005;121:1447-8.

5. Alanen P. Näyttö ja lääketiede. Duodecim 1999;115:2437-41.

6. Poikolainen K. Näyttöön perustuva lääketiede: hullunkurinen ilmiö? Suom Lääkäril 2002;57:4853.

7. Huittinen VM. Näyttöön perustuva lääketiede ja toinen totuus. Suom Lääkäril 2002;57:5099.

8. Leiman M. Vaikuttavuustutkimuksen pulmallisuus psykoterapiassa. Duodecim 2004;120:2645-53.

9. Leiman M. Psykoterapiamuodot ja käypä hoito. Duodecim 2006;122:701-2.

10. Keinänen M, Lahti I, Pylkkänen K. Psykoterapia ja näyttöön perustuva lääketiede: mielen vuorovaikutuksellinen tieteenteoria. Suom Lääkäril 2006;61:1221-7.

11. Louhiala P, Hemilä H. Näyttöön perustuva lääketiede: hyvä renki mutta huono isäntä. Duodecim 2006;121:1317-25 and 1448-9.

12. Hemilä H. Lääketieteessä on monenlaisia näyttöjä. Suom Lääkäril 2011;66:3111.

13. Medline data base (http://www.ncbi.nlm.nih.gov/pubmed) searches were carried out through the Ovid Gateway in August 2012.

14. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7:e1000326.

15. Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. J Clin Epidemiol 2000;53:1193-9.

16. Prins JM, Büller HR. Meta-analysis: the final answer, or even more confusion? Lancet 1996;348:199.

17. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized controlled trials. N Engl J Med 1997;337:536-42.

18. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol 2009;9:10.

19. Hemilä H. Some examples of problems with meta-analysis. Available at: http://www.mv.helsinki.fi/home/hemila/metaanalysis/problems.htm

20. Singh M, Das RR. Zinc for the common cold. Cochrane Database Syst Rev 2011;(2):CD001364.

21. Hemilä H. The zinc for the common cold review by Singh and Das has a number of problems [Feedback]. In: Singh M, Das RR. Zinc for the common cold. Cochrane Database Syst Rev 2011;(2):CD001364. Available at: http://www.mv.helsinki.fi/home/hemila/H32P.pdf

22. Hemilä H, Kaprio J. Modification of the effect of vitamin E supplementation on the mortality of male smokers by age and dietary vitamin C. Am J Epidemiol 2009;169:946-53.

23. Vandenbroucke JP. Why do the results of randomised and observational studies differ? BMJ 2011;343:d7020.

24. Hróbjartsson A, Gøtzsche PC. Placebo interventions for all clinical conditions. Cochrane Database Syst Rev 2010;(1): CD003974.

25. Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? CMAJ 2006;174:645-6.

26. Vandenbroucke JP. Case reports in an evidence-based world. J R Soc Med 1999;92:159-63.

13th Physiatric Summer School

PLACEBO

Placebo Effects in Musculoskeletal Disorders

Editor: Karl-August Lindgren

16. 8. - 17. 8. 2012 Helsinki

REHABILITATION
ORTON
INVALID FOUNDATION

Published by
Rehabilitation ORTON
ORTON Foundation
Tenholantie 10
00280 Helsinki

Design by
Tommy Kaj Lindgren

Printed at
Vammalan Kirjapaino,
Vammala 2013

ISSN 1796-4148
ISBN 978-952-9657-67-4

13th Physiatric Summer School

Placebo
Placebo Effects in Musculosceletal Disorders

Editor: Karl-August Lindgren