REVIEW

Data *and* models determine treatment proposals—an illustration from meta-analysis

U Helfenstein

Postgrad Med J 2002;**78**:131–134

A relevant problem in meta-analysis concerns the possible heterogeneity between trial results. If a test of heterogeneity is not significant the trials are often considered to be "homogeneous" and the individual trial results are replaced by an overall mean effect size and its confidence interval ("equal effects model"). If the trials are heterogeneous the individual trial effect sizes are conserved ("fixed effects model"). In a more flexible approach ("random effects model"), each trial makes use of knowledge from the other trials so individual effect sizes are "shrunken" towards an overall mean effect size. The more flexible tool may be useful for doctors involved in a trial when the outcome of their individual trial differs markedly from the overall mean effect size. Where a particular trial result is opposite in direction to the overall mean result, a conflict may arise: should a new patient be treated with the new method or not? The more flexible position and a graphical comparison of the three approaches are likely to be helpful in guiding the decision. Applying different models to the same data may lead to apparently paradoxical results: an individual trial result may be interpreted to be beneficial or harmful depending on the choice of model.

Correspondence to:
Dr Ulrich Helfenstein, Department of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Sumatrastrasse 30, 8006 Zurich, Switzerland

Submitted 25 June 2001
Accepted 31 August 2001

It is now widely recognised that meta-analysis may be of vital importance in order to analyse, interpret, and communicate a collection of clinical trials. For the last few years, meta-analysis has been of increasing interest in medicine. This is reflected, for example, in consecutive series of tutorial articles devoted to this topic.[1–6] The topic has also been taken up in modern introductory books on medical statistics.[7]

Following the terminology of Laird and Mosteller three basic statistical models are considered in meta-analysis[8]:

(1) Fixed effects model (FEM).

(2) Equal effects model (EEM).

(3) Random effects model (REM).

They differ in particular with regard to "pooling" the effect sizes obtained from the individual studies into an "overall effect size". The FEM assumes that the differences between the studies are so important that pooling is not indicated and that individual effect sizes should be retained. If heterogeneity between the studies has been established, it is natural to choose the FEM. In this case a mean effect size would pertain to a somewhat hypothetical population.

At the other extreme, the EEM assumes that the true unknown effect sizes of all trials in the meta-analysis are equal. All observed variation is due to sampling or within trial variability. The estimates of the individual effect sizes are replaced by the estimate of the overall effect size. In these two models it is assumed that the effect sizes are fixed unknown quantities; therefore the EEM is often also called a "fixed effects model".

In strong contrast, in the REM it is assumed that the individual trial effect sizes are "random" quantities. Other teams at other places in the world could also have performed similar studies. The reported trials are considered to be a random selection of trials from a conceptual population of trials which has a more or less pronounced between trial variation.

It is well known that the EEM and the REM approaches differ with regard to the estimation of the overall effect size and its confidence interval. The REM usually provides broader confidence intervals allowing for between study variation. The EEM gives narrower confidence intervals or "more significant" overall effect sizes. This may explain that the EEM enjoys more popularity. In addition, the EEM and the REM differ in weighting the effect sizes of individual trials when calculating the overall effect size. Estimates of overall effect sizes and standard errors for both models can be calculated conveniently by applying freely available software.[9]

In this outline, interest concentrates not on overall effect sizes and corresponding confidence intervals, but rather on the estimates of the individual trial effect sizes obtained by the different models.

In a "classical" randomised clinical trial with a treatment and a control group each trial provides four numbers: the sample sizes in the control group and in the treatment group and the number of events, for example deaths, in each of the two groups. A well established measure of effect size is the odds ratio (OR) or its logarithm (logOR).

For the sake of illustration, assume that a series of such trials has been conducted in different clinics X, Y, Z, ... A meta-analysis has been performed and presents a list of ORs: OR_x, OR_y, OR_z, ... together with an overall mean estimate

Abbreviations: EEM, equal effects model; FEM, fixed effects model; logOR, logarithm of odds ratio; REM, random effects model

$OR_{overall}$ and its confidence interval. What should a doctor in clinic X etc learn from the meta-analysis and how should he treat his next patient?

Accepting the EEM implies that he may "forget" the estimate in his own clinic OR_x and replace it by the overall estimate $OR_{overall}$. In an extreme case, the study in clinic X could have shown a harmful treatment effect OR_x. In contrast, the $OR_{overall}$ could have shown a significant beneficial treatment effect. Following the interpretation of the EEM the doctor in clinic X should use the treatment, even though the study in his clinic showed a harmful treatment effect.

In another situation a meta-analysis detects pronounced between study variation (for example by using a Cochran test of heterogeneity) and chooses the FEM to summarise the trials. What should a doctor in clinic X etc learn in this case from the meta-analysis? He should essentially stay with the OR_x resulting from the trial in his specific clinic X. In the above extreme situation, he should not use the treatment found to be harmful for his patients!

The decision between the models may be delicate. If a test of heterogeneity of trials is not significant, it is often decided to apply the EEM. This model has the attractive looking but possibly misleading property that it usually gives narrow confidence limits for $OR_{overall}$. However, a non-significant test does not mean that the effect sizes are equal. In addition, the test may not be powerful to detect heterogeneity of trials.[2]

The EEM and the FEM represent idealised and extreme situations. It is natural to think that a "compromise" between the two extreme positions approximates the unknown truth better: the more flexible REM[2] allows us to obtain this compromise. Instead of presenting its intricate statistical structure, the following example should, by way of illustration, help to develop some intuition for its practical performance.

An illustrative example

In order to illustrate the differences between the three approaches a dataset of 22 trials published by Yusuf *et al* (table 10, page 349) is used.[10] The trials were concerned with a comparison of mortality after myocardial infarction between treated (prophylactic use of β-blockers) and control groups. The dataset has been used to describe mathematical statistical aspects of meta-analysis.[11 12] It is thought that this dataset is also well suited to demonstrate basic differences between the three approaches in a non-technical manner and to develop some intuition for them.

Figure 1 shows the results (logORs) of the three models concerning the estimates of the individual trials. The logORs pertaining to the same trial are connected by a line. A horizontal line separates "pessimistic" trials with a positive logOR (that is OR >1, harmful result) from "optimistic" trials with a negative logOR (that is OR <1, beneficial result). On the left side (A) the individual logORs are depicted corresponding to the FEM. On the right side (D) the logOR arising from the EEM is shown. In the middle panel (C) the estimates of the individual logORs as obtained from the REM are presented. The differences are quite impressive.

Assume for a moment that you are a doctor involved in conducting trial 14 (OR_{14} = 1.33; $logOR_{14}$ = 0.282). What can you learn from the "three model graph" in fig 1 with regard to the substantial meaning of the models?

If you decide to accept the FEM shown in the left panel (A) you might conclude that for patients entering your clinic you should essentially stay with the OR_{14} resulting from the trial in your specific clinic. In the above extreme situation, you might perhaps not use the treatment found to be *"harmful"* for your patients.

In contrast, if you decide to accept the EEM shown on the right side (D) you have to "forget" the outcome of your own trial OR_{14} and replace it by the overall estimate $OR_{overall}$. $OR_{overall}$ shows a beneficial treatment effect. Following the interpretation of the EEM you have to *use* the treatment *even though* your own trial showed, if anything, a *"harmful"* treatment effect for your patients. This paradoxical situation may be difficult to accept even if the individual trial effect is not significant.

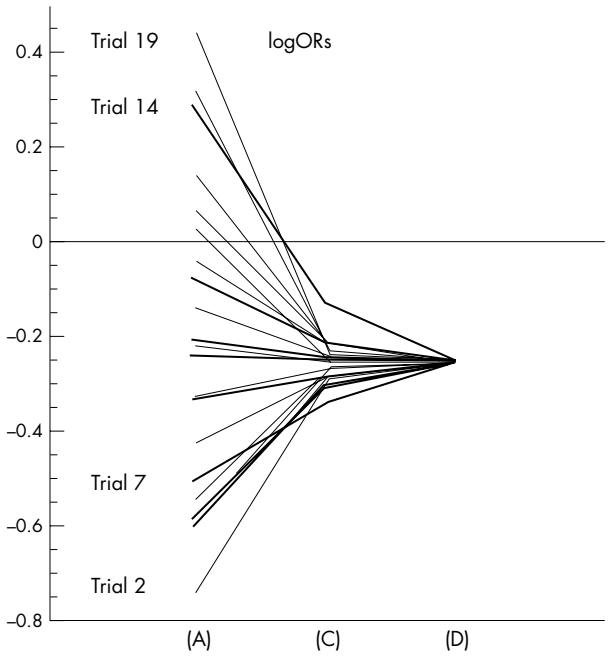

Figure 1 Three approaches to the estimation of the individual trial effect sizes (logORs). (A) Fixed effects model (FEM), "original" trial effect sizes. (C) "Shrunken" estimates of trial effect sizes as obtained from the random effects model (REM). (D) Equal effects model (EEM): one effect size for all trials. For explanation see text.

Analogously, assume you are a doctor involved in conducting trial 2 (OR_2 = 0.477; logOR2 = –0.741). If you accept the FEM shown on the left (A), you might conclude that the treatment is most successful for a typical patient coming to your clinic. Following the EEM on right side (D) you have to revise quite drastically this "overoptimistic" view. It is likely that you will not be happy to learn that you may forget your particular success.

Thus, the left side (A) and right side (D) visualise extreme positions. With regard to treatment of patients from your own clinic you can, loosely speaking, "forget the other trials" on the left side and "forget your own trial" on the right side. It is likely that a "compromise" is more realistic. This compromise as obtained by means of the REM is depicted in the middle panel (C). It demonstrates how the result of your own trial is *modified* by knowledge of the outcomes of the other trials in the meta-analysis. The trials, particularly the extreme trials, are now "shifted" or "shrunken" towards a mean effect.

On the left side, six of the 22 trials are on the "pessimistic side"; they show a logOR >0 (that is an OR >1, an increased number of deaths in the treatment groups). In the middle panel *all* these trials are shifted to the "optimistic side" where logOR <0 (decreased number of deaths in the treatment groups). Analogously, trials showing strong beneficial effect of treatment are shifted upwards to a less optimistic view. Trialist 7, for example, has to modify his optimistic view. However, the corresponding patients appear to benefit still more from treatment than suggested by the right panel (EEM). Trials showing a medium beneficial effect are not shifted markedly when comparing the left and the middle panel.

The degree of shift towards the overall logOR varies from trial to trial. In particular, an interesting observation, surprising at first sight, is the "crossing of lines". A trial with a "pessimistic view" in the left panel (for example trial 19 with the

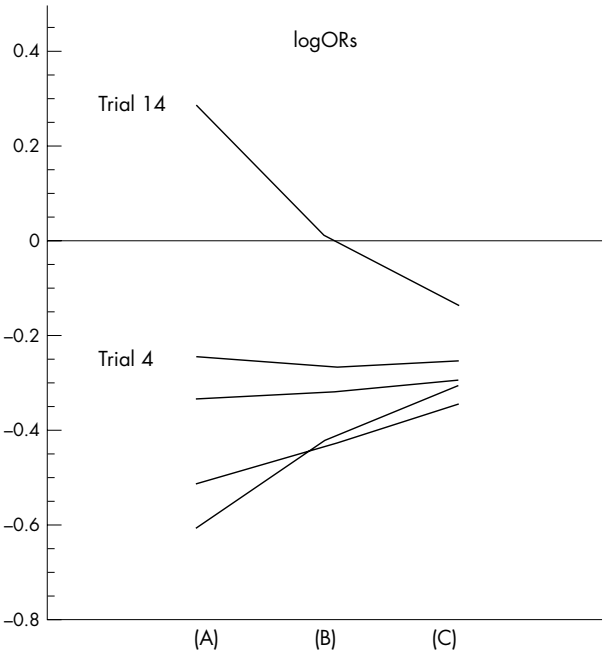

Figure 2 (A) "Original" trial effect sizes (same values as the corresponding in (A) of fig 1). (B) REM estimates of trial effect sizes calculated from the subset of the five selected studies only. (C) REM estimates of trial effect sizes calculated from all 22 studies. For explanation see text.

"worst" result, logOR = 0.444) may become more "optimistic" in the middle panel than a less pessimistic trial (for example trial 14 with logOR = 0.282). This effect is explained by the smaller sample size of trial 19 (with n = 308). Trial 19 has to "learn" more from the other trials. Trial 14 (with n = 1741) contains more information with regard to patients entering the corresponding clinic. It has to learn less; it is shifted less. The effect of sample size is also graphically demonstrated by using bold lines for the larger trials (sample size larger than 750). The larger trials are clearly shrunken less than the smaller trials.

In order to illustrate the influence of the presence or absence of other trials on the REM estimates of individual trial effects a subset of five large trials (trials 4, 7, 10, 14, and 22) has been selected from the 22 trials and a separate analysis has been performed. Figure 2 shows the original individual trial effect sizes (logORs) as obtained by the FEM and the estimates of the effect sizes obtained from the two REMs; one including all 22 trials and the other including only the five selected trials. For brevity the result of the overall effect size corresponding to the EEM has been omitted in fig 2.

The FEM including only five trials provides as estimates the original, individual logORs. They are the same as those obtained by inclusion of all 22 trials and they are shown on the left side (A). On the right side (C) the REM estimates calculated from all 22 studies are shown. These are the same values as the corresponding five values contained in the middle panel (C) of fig 1. In the middle panel (B) the REM estimates calculated from the subset of the five selected studies only are shown.

The differences of the estimates in fig 2 between the middle and right side are not due to sample size (the studies are the same). They are due to the influence of the results of the other studies. The shrinkage from left to right (22 trials) is larger than from left to middle (five trials). The REM estimates shown in the middle panel differ from those shown on the right. The "shrinkage effect" is less pronounced in the five trials analysis than in the 22 trials analysis. Each trial "learns more" when taking into account all the 21 remaining trials than when taking only four others into account.

Consider in particular trial 14. It exhibits the "worst" result in all three approaches. However, in the five trials REM (B) the estimate just stays on the "pessimistic side" while in the 22 trials REM (C) it changes to the "optimistic" side. Thus, even when only REMs are considered, the same trial may appear on the optimistic side or on the pessimistic side depending on the information contained in the other trials.

In model (B), four of the five selected trials show a logOR which is intermediate between the logORs of model (A) and model (C). In contrast, trial 4 shows a slightly smaller logOR in (B) than in (A) and (C). This shift may be explained by the fact that in (B) three of the trials have a smaller effect size and only one trial has a larger effect than trial 4.

DISCUSSION

Meta-analysis involves many aspects and problems. In the present article one topic is concentrated on: how should one modify the interpretation of any particular trial outcome in the light of sole results of the other trials and how should we treat accordingly our next patient?

Choosing the EEM signifies that the estimates of the individual effect sizes are replaced by the estimate of the overall effect size. In contrast, the FEM conserves individual effect sizes when considering individual trials. In short: EEM = "full shrinkage" ; FEM = "no shrinkage".

In the REM, estimates of individual effect sizes are shrunken in an intermediate manner. They are shifted more or less to the overall mean effect size where the degree of the shift depends in particular on the sample sizes. Each trial makes adequate use from knowledge of the other trials.

It is likely that such a flexible "compromise" may be more appropriate than the two extreme and rigid approaches. Doctors may find themselves in a severe conflict when the effect size of their own trial is opposite to the overall effect size as proposed by a EEM. It is obvious that they will feel more convinced of a meta-analysis if they can see how much of the information of their own trial result is retained. Using a REM they may see that they have not to forget their particular trial outcome; each particular effect size is "only modified" more or less by the additional knowledge available from the other trials. In particular, a graphical representation of the individual effect size estimates corresponding to the three approaches, such as the one presented in fig 1, helps a trialist to clarify where his trial stays in the collection of all trials. Figure 1 should be particularly helpful to develop some intuition for the essential meaning of the three models.

If I am one of the doctors in clinic X which conducted trial X and I know only that my new patient belongs to the same "input stream of patients" of which the patients of trial X were taken, it is natural to adopt the modified result of the REM and to treat the patient accordingly. In contrast, if I have no prior knowledge to which of the input streams of the trials my new patient is closest, I expect a treatment effect as proposed by an overall mean effect size.

In the present illustrating outline several relevant topics and the technical details have been omitted. For example, if additional knowledge about *covariables* such as average patient age of each trial is available, it may be of interest to include this knowledge in *more complex* models. Sometimes it may be possible to perform "subgroup analysis" to explore sources of heterogeneity. For clarity, the present outline concentrates on the three basic models considered to be of prime relevance.

Thus, some suggestions to the more in depth and comprehensive literature may be helpful. The description of the EEM and the FEM with the corresponding formulas may be found in the manual of EasyMA.[9] In the above example a so called "Bayesian" REM was estimated. A deeper insight into this approach may be found in Carlin.[11] A reader interested in applying the very recommendable and freely available

software, BUGS, finds a very valuable discussion of the Bayesian approach to estimation of complex models in its manual.[12 13] In addition, the modern methods of estimation via Monte Carlo simulation (Gibbs sampling[14], Markov Chain Monte Carlo or MCMC[15]) are clearly presented. A very helpful tool to visualise complex hierarchical models such as the Bayesian REM, the so called "conditional independence graph", is carefully explained. In addition, the application of the software is demonstrated by presenting many worked examples including meta-analysis.

The flexible "intermediate" model above has, for simplicity, be called a REM. Two variants are used mainly. In addition to the Bayesian approach, there is an other related model providing similar estimates which is called "empirical Bayesian" REM.[16 17] Both, the "Bayesian" and the "empirical Bayesian" have their advantages. For example, the Bayesian method gives somewhat more realistic, larger standard errors than the empirical Bayesian method. The intricate statistical problems involved prohibit a discussion here. It is thought that at the present time the carefully documented, very general, flexible, and freely available software BUGS[12] provides a strong argument for performing analyses using the Bayesian approach. In addition, if information about covariables is available, the Bayesian approach using BUGS allows conveniently to include it in a more complex model. Thus, when conducting a meta-analysis it is recommendable to complement the usual analysis where only one of the two extreme and rigid models is performed by the flexible REM using BUGS and by a graphical representation of the trial effect size estimates corresponding to the three models.

The example demonstrates that applying different models to the same data may lead to apparently paradoxical results: an individual trial result may be interpreted to be beneficial or harmful depending on the choice of model. This illustrates that models are not only of theoretical interest but of basic practical relevance.

Key points

- The choice of model may be of vital relevance; it should not just be fully delegated to the statistician.
- In the context of meta-analysis there are essentially three models of choice: (1) the "equal effects model", (2) the "fixed effects model", and (3) the "random effects model".
- Applying the different models to the same data may lead to apparently paradoxical results: an individual trial result may be interpreted to be beneficial or harmful depending on this choice of model.
- It is recommendable to develop some intuition for the substantive meaning of models.

REFERENCES

1 **Egger M**, Davey Smith G. Meta-analysis: potentials and promise. *BMJ* 1997;**315**:1371–4.
2 **Egger M**, Davey Smith G, Phillips AN. Meta-analysis: principles and procedures. *BMJ* 1997;**315**:1533–7.
3 **Davey Smith G**, Egger M, Phillips AN. Meta-analysis: beyond the grand mean. *BMJ* 1997;**315**:1610–4.
4 **Egger M**, Davey Smith G. Meta-analysis: bias in location and selection of studies. *BMJ* 1998;**316**:61–6.
5 **Egger M**, Schneider M, Davey Smith G. Meta-analysis: spurious precision? Meta-analysis of observational studies. *BMJ* 1998;**316**:140–4.
6 **Davey Smith G**, Egger M. Meta-analysis: unresolved issues and future developments. *BMJ* 1998;**316**:221–5.
7 **Altman DG**. *Practical statistics for medical research*. London: Chapman and Hall, 1991.
8 **Laird NM**, Mosteller F. Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care* 1990;**6**:5–30.
9 **Cucherat M**, Boissel JP, Leizorovicz A. EasyMA: a program for the meta-analysis of clinical trials. *Computer Methods and Programs in Biomedicine* 1997;**53**:187–90.
10 **Yusuf S**, Peto R, Lewis J, *et al*. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progr Cardiovasc Dis* 1985;**27**:335–71.
11 **Carlin**, J B. Meta-analysis for 2×2 tables: a Bayesian approach. *Stat Med* 1992;**11**:141–58.
12 **Spiegelhalter DJ**, Thomas A, Best NG, *et al*. *BUGS: Bayesian inference using Gibbs sampling, version 0.50*. Cambridge: MRC Biostatistics Unit, 1997.
13 **Gilks WR**, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. *The Statistician* 1994;**43**:169–78.
14 **Casella G**, George EI. Explaining the Gibbs sampler. *American Statistician* 1992;**46**:167–74.
15 **Gilks WR**, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in practice*. London: Chapman and Hall, 1996
16 **Casella G**. An introduction to empirical Bayes data analysis. *American Statistician* 1985;**39**:83–7.
17 **Louis AT**. Using empirical Bayes methods in biopharmaceutical research. *Stat Med* 1991;**10**:811–29.