A look at the statistical overview (or meta–analysis)

ABSTRACT—In essence, the statistical overview (or meta-analysis) does little more than many medical scientists and practising clinicians have been attempting for years: estimating the average size of an effect from an apparently disparate medical literature. The statistical overview provides a method of pooling the results of comparable trials (or observational studies) which, on their own, may be too small to demonstrate a 'statistically significant' effect. The role for the overview is obvious, the method is relatively simple, but the findings have to be interpreted cautiously, since joining the hypothesis post hoc contravenes one of the essential principles of scientific enquiry. The use of the statistical overview and some of its pros and cons are illustrated in four clinical examples.

The prima facie role of the statistical overview is clear. By pooling the results of small trials or the findings of small studies the reviewer may obtain an estimate of the average effect of an intervention or the average magnitude of an association. These are logical steps in scientific inquiry in the development from the initial observation (the n = 1 trial), through small controlled experiments to large empirical population-based trials. In the present age of mega-trials and of biostatisticians exhorting increased power, some forget the importance of the original (single) observation and some criticise early studies on their inability to distinguish quite large and possibly clinically important differences. Each has its place in the advancement of knowledge. Like the researching scientist, the (practising) clinician needs to know the effect of an intervention or therapy or the magnitude of an association in the light of current knowledge. Therein lies the essential role for competent overview.

In the past, reviews were most often written as editorials (often anonymously) by the (senior) authority in the field, calling on his accumulated experience of the relevant published or unpublished research. While the author of the editorial was anonymous, the statistician was even more of a background figure. Although it may be popularly believed that assessment of evidence was judged 'by experience', statistical advice was available and was sought and acknowledged in planning studies and in reporting results. The ‘explosion’ in the second half of the century in scientific research, medical therapies, clinical trials evaluating therapies, and medical journals reporting trials has led to formalisation of assessment of scientific (or medical) evidence [1,2] and to a shift in emphasis from the art of writing a review to the science of reviewing the evidence. Since clinicians and medical scientists seldom deal with equations and certainty, but usually with multifactorial aetiology and biological variation in the therapeutic response, the findings of each individual study and of each individual trial remain imprecise predictors of the hypothetical ‘true’ effect. The statistical overview (or meta-analysis) is intended to improve prediction of the estimated effect.

Statistical overviews have been rapidly introduced in recent years and may now be seen in many clinical journals. Some of the pros and cons of the technique are discussed here around four examples, some of which may be regarded as among the current 'controversies in medicine'. The reader seeking more technical description and criticism should refer to papers mostly in the statistical or biostatistical journals (references later).

Method (principle)

In essence, the statistical overview pools the findings or results of comparable studies or trials to obtain an overall average. In effect, the statistician coalesces a number of small trials into one (powerful) trial. For example, if five of six small trials give 'non-significant' results, pooling six trial conclusions would suggest that overall the effect was 'non-significant'. However, if each trial reported five-year survival benefits (of a new drug) of, say, 10% over the control, pooling five-year survivals of patients in all six trials would suggest a benefit of approximately 10%, which might well achieve 'statistical significance' at p < 0.05. This is part of the rationale for the move from hypothesis testing (p values) to estimation of relative risks (RR), or odds ratios (OR) and confidence intervals (CI). For a single trial a clinician's interpretation is not very different: viz, the relative risk of death within five years (on the new drug compared with the old) is estimated as 0.90 (with 95% confidence interval of say 0.72-1.11). The confidence interval embraces 1.00 (evens), so the clinician's interpretation is that the relative risk of death is 'not significant' (not significantly different from 1.00). The estimation approach, recommended by medical statisticians [3], certainly aids conceptualisation of the

Robert R West, PhD, FSS, MFPHM (Hon)
Reader in Epidemiology, University of Wales College of Medicine, Cardiff
Effect of adding trials together: e.g., six trials each with RR = 0.90 would improve the precision of the estimate of relative risk, and the 95% confidence interval of the pooled results would narrow to, say, 0.84–0.96, no longer embracing 1.00 and therefore becoming ‘significant’. (Note: just as with p values in hypothesis testing, the confidence interval is only a guide; the true answer does not lie ‘within’ or ‘not within’ the confidence interval.)

Effect of publication bias

The statistical overview effectively treats a number of individual trials as though they were one large trial, adding the number of events (deaths) in all trials on the new drug/intervention side and comparing them with the total number of events on the placebo/control side. There are certain requirements for the pooling exercises to be valid [4,5]. The trials should test the same hypothesis, comparing similar patients (diagnosis, age, sex, stage of disease, comorbidity), or similar interventions (drugs, doses, operations), and measuring the same outcome or endpoint (survival, wellbeing, physiological function). All trials should be considered, whether published or not, and the inclusion/exclusion criteria should be chosen without knowledge of the effect the trial has on the pooled outcome. These latter considerations are impracticable and consequently the statistical overview, by its retrospective nature, contradicts one of the principles of scientific enquiry: first the hypothesis, then the test. The overview starts with a number of answers and then selects from those answers a sample that satisfies certain criteria. Guidance on selecting trials on quality may be found elsewhere [1,2,5,6]. Quality may be employed simply to decide selection or rejection or may be used as a score to weight the evidence contributed by each selected trial. Thorough overviews may seek extra information from original authors not reported in the selected papers [4]. The possible effect of publication bias cannot be overstated: publication bias in its fullest sense includes investigator bias, investigation bias, trial completion bias, writing up and submission bias, editorial bias, referee bias, and revision bias (following advice of editors and referees). The effect of publication bias is well illustrated in cancer chemotherapy trials: because such trials are usually registered, the pooled benefit shown by all registered trials may be equivocal, yet when restricted to published trials, the pooled benefit may be judged highly significant [7]. Several of these potential hazards in pooling the results of published trials are illustrated in the following examples.

Example 1 (rehabilitation following myocardial infarction)

In a trial of rehabilitation following myocardial infarction by physical training, the three-year mortality was summarised as follows: exercise training 12/151 died and control 21/152 died [8]. On its own this trial does not demonstrate that rehabilitation by exercise training increases three-year survival at the conventional level of significance (p < 0.05). While the trial was of admirable size at the then state of knowledge and adequate to test possibly bigger effects on such physiological variables as oxygen uptake, exercise tolerance, and work capacity, it had insufficient power to demonstrate a possible clinically important effect on survival, because there were too few deaths. Nevertheless, although ‘not significant’, the trial estimates a large (possible) improvement in survival (or reduction in three-year mortality) with RR = 0.58 (and CI = 0.29–1.13). If other similar trials give similar results (each perhaps individually ‘non-significant’), this relatively large and potentially important clinical effect might be statistically significant. It is, therefore, of interest to apply the techniques of the statistical overview to estimate the extent of the reduction in mortality in the pooled data.

The principal findings of a typical statistical overview are illustrated in Figure 1, derived from data of the WHO collaborative trial [9]. The presentation follows current statistical recommendations, the short vertical line representing the point estimates of relative risk (of each trial) and the horizontal whisker representing its 95% confidence interval. Estimates to the left of the central line represent trials which suggest that benefit was conferred by rehabilitation programmes. Relative risks of death at three years are shown on a logarithmic scale, since that represents

| Study | Patients |
|-------|----------|
| Sato  | 33       |
| Priker| 63       |
| Marie | 36       |
| Campos| 150      |
| Gruber| 34       |
| Zhou  | 51       |
| Wood  | 50       |

Total 417

Relative risk (log scale)

0.1 0.2 0.5 1.0 2.0

Fig 1. Statistical overview of 17 trials of rehabilitation following myocardial infarction: relative risk (and 95% confidence intervals) of mortality at three years comparing rehabilitation programme patients with controls. (Ref 9).
comparable benefits to rehabilitation or control as equidistant on their respective sides of the central line of evens (relative risk of death at three years = 1.0). Individually, only two centres gave 'significant' results: Moscow better and Ghent worse.

One published overview included 14 centres from the WHO collaboration and seven independent trials [10]. The two groups are not strictly comparable, since six of the independent trials were described by their authors as exercise training programmes, whereas the WHO trial was of multifactorial rehabilitation, including an exercise training component. It is interesting to compare the original report of the WHO collaborative group (two of 17 centres significant, one better and one worse [9], which was interpreted by most clinicians as 'equivocal'), with the findings in this overview of 14 centres selected from the WHO collaboration (RR = 0.79), 95% CI 0.62–1.01), which becomes 'significant' when the pooled exercise only trials (with similar RR) are added [10]. O'Connor et al give reasons for excluding three centres reported by the WHO collaborative group, who incidentally themselves excluded seven centres of the original collaboration [9]. This example illustrates the potential for the statistical overview. Hypothesis testing in 17 centres suggested that multidisciplinary rehabilitation was unlikely to improve survival, while estimation of the average benefit among the same 17 centres (or 2,600 patients) suggested a relative risk of mortality almost 'significantly' down by 18%.

Example 2 (heparin following surgery)

Low molecular weight (LMW) heparin has been compared with unfractionated heparin for its effect on several post-surgery complications in many randomised controlled trials. The incidences of deep vein thrombosis (DVT) and pulmonary embolism (PE) following general surgery are not high, and individually most trials showed no significant benefit of LMW heparin. Recently, two statistical overviews have reported pooled estimates of the relative risks of these complications with slightly different conclusions [11,12]. For example, for DVT (Table 1) the former overview selected 23 trials (of 43 identified) and reported a relative risk of 0.74 (0.65–0.86) but went on to conclude that there was no convincing evidence for LMW heparin in general surgery on the basis of eight general surgery trials selected for methodological rigour, RR = 0.91 (0.68–1.23). On the other hand, the latter overview selected 39 trials and reported an overall relative risk of 0.85 (0.74–0.97) and concluded, with this apparently less significant pooled estimate, that LMW heparin was more efficacious than unfractionated heparin in general surgery on the basis of the risk in all 25 general surgery trials being of the same order RR = 0.86 (0.72–1.04), although not 'significant' by themselves.

It is interesting that two groups, having undertaken statistical overviews of the same trial literature within weeks of each other, obtained rather different overall relative risk (0.74 v 0.85) and then drew different conclusions with respect to efficacy in general surgery, although their pooled estimates of the relative risk for general surgery were closer (0.91 v 0.86) [13]. Without becoming too involved in the rights and wrongs of either study or of LMW heparin, these differences illustrate how trial selection influences the final estimate of effect, suggest that interpretation of the meaning or 'realness' of (final estimated) effect often depends on considerations outwith the overview, and possibly suggest that anticipation of the 'real' effect may influence trial selection.

Example 3 (cholesterol lowering in populations)

Another example that deserves consideration, as a leading current controversy, is that of the cholesterol-lowering trials. Some 24 major trials of cholesterol lowering by diet or drug have reported all-cause mortality. Six statistical overviews have been reported [14–19], and no doubt many others have been completed although not formally reported [20–23]. The estimated relative risk of death in the cholesterol-lowering groups compared with controls, according to the pooled findings of the six published overviews are summarised in Figure 2. In all overviews, the pooled relative risk of all-cause death is near unity and the 95% confidence interval encompasses unity. The two smallest poolings give the most widely separated estimates (0.92 and 1.08) and the two largest poolings give estimates near 1.0.

The correspondence following publication of some of the more recent overviews has been illuminating [24]. Some authors were criticised for being selective, choosing trials that suited their purpose, while others were criticised for being too lax in their selection criteria, although the findings of all support the view that the net effect on mortality is nil, or at best equivocal. Ravnskov observed that reviewers (statistical overviews) who were not seeking a benefit in lipid lowering found no more favourable result in included trials than in excluded trials, while reviewers who consid-

| Overview                        | Trials included | Relative risk of DVT (95% confidence interval) |
|--------------------------------|----------------|-----------------------------------------------|
| Nurmohamed et al (ref 11)       | 23             | 0.74 (0.65–0.86)                              |
| General surgery                 | 8              | 0.91 (0.68–1.23)                              |
| Leizorovicz et al (ref 12)      | 39             | 0.85 (0.74–0.97)                              |
| General surgery                 | 25             | 0.86 (0.72–1.04)                              |
Considered lipid lowering to be a ‘good thing’ found more favourable relative risks in trials selected [1,3,4 in Fig. 3] than in those excluded (1.11,1.14, and 1.15 respectively), which hints at selective inclusion [19]. The conclusion, nevertheless, is inescapable, that despite possible selection bias or possibly over lax selection criteria, all six selections from the published literature point towards no net improvement in mortality from the lipid-lowering regimes evaluated to date. Within an unchanged total mortality, many trials (and all overviews) suggest a shift in cause from cardiovascular to cancer and ‘non-natural’, varying somewhat in magnitude from one trial (or overview) to another.

It is interesting to consider to what extent the rather disappointing finding in these lipid-lowering trials could have been anticipated by study of all-cause mortality in the many observational cohort studies that used cholesterol as a ‘risk marker’ [23]. The total population included in cohort studies (approx 32 studies and 650,000 population) greatly exceeds the population included in lipid-lowering trials referred to above (24 trials and 45,000 patients). While statistical overview (meta-analysis) has mostly been used for pooling the results of similar experimental trials, the technique is applicable also to pooling the findings of observational studies, again provided that the studies are comparable, test a common hypothesis, include similar populations, compare the same risk factor (or marker) and comparable exposure (or dose), and compare the same outcome. There is more heterogeneity between cholesterol cohorts and any pooling is likely to be heavily weighted by one single large cohort study (the MRFIT study), but individually all cohort studies show a shallow U-shaped curve or a near flat relationship between cholesterol and total mortality and a transition with increasing levels of cholesterol from high cancer/low cardiovascular to low cancer/high cardiovascular mortality [25].

**Example 4 (multi-drug resistant phenotype and remission)**

The potential for the statistical overview in observational study may be illustrated in a typical example, in
which medical scientists and clinicians might appreciate the value of the statistical overview (perhaps without calling on a statistician to estimate the pooled odds ratio). Ten studies have examined the relationship between the multi-drug resistant (MDR) phenotype and remission or survival in acute myeloblastic leukaemia [26]. As so often in the early stages of a developing medical literature, many studies are small and individually may not achieve ‘significance’. Pooling seven studies that reported remission rates provides an estimate of the relative risk = 0.52 (0.44–0.62). (Fig 3). Because of the influence of publication bias, it is interesting also to consider the effect of including and estimating for two larger studies, reported only in abstracts as ‘not associated’. Allowing for these two unpublished studies, the final estimate is that the relative risk remains significantly reduced [26].

Conclusion

In conclusion, the statistical overview (or meta-analysis) is not really such a foreign beast. It really only formalises what clinicians have been doing for years, estimating an average effect from currently available reports. The principal argument in its favour lies in the fact that several small trials (or observational studies) may be individually ‘not significant’ yet may mask an important clinical effect, which becomes clear (and ‘significant’) in the pooled data. It should not be mistaken for a large primary trial (or study). But there are important arguments for caution in its application, particularly in respect of publication bias (in its fullest sense) and the influence of inclusion/exclusion criteria being inescapably linked with the overviewer’s prior prejudice, since by its very nature overviewing is all post hoc [5,7,24]. Because there are so many publications, the statistical overview is here to stay. Medical scientists and clinicians need to appreciate its potential uses and limitations, and to interpret it cautiously.

References

1 Gehlbach SH. Interpreting the medical literature. Circulation 1990;82:1916–24.
2 Elwood JM. Causal relationships in medicine. London: Oxford Medical, 1988.
3 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986;292:746–50.
4 Peto R. Clinical trial methodology. Biomedicine (special issue) 1978;28:24–36.
5 Thompson SG, Pocock SJ. Can meta-analysis be trusted? Lancet 1991;337:1127–30.
6 Chalmers TC, Smith H, Blackburn B, et al. Method for assessing quality of a randomised controlled trial. Controlled Clin Trials 1981;2:31–49.
7 Begg CB, Berlin JA. Publication bias, a problem in interpreting medical data. J R Stat Soc 1988;151:419–63.
8 Carson P, Phillips R, Lloyd M, et al. Exercise after myocardial infarction: a controlled trial. J Roy Coll Phys 1982;16:141–7.
9 World Health Organisation. Rehabilitation and comprehensive secondary prevention after acute myocardial infarction. EURO report 84, Copenhagen: WHO, 1983.
10 O’Connor GT, Buring JE, Yusuf S, et al. Overview of randomised trials of rehabilitation after myocardial infarction. Circulation 1989;80:234–44.
11 Nurmohamed MT, Rosendaal FR, Buller HR, et al. Low molecular weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. Lancet 1992;340:152–6.
12 Leizorovicz A, Haugh MC, Chapuis FE, et al. Low molecular weight heparin in the prevention of perioperative thrombosis. Br Med J 1992;305:913–20.
13 Routledge PA, West RR. Low molecular weight heparin. Br Med J 1992;305:906.
14 Lipid Research Clinics Program. The lipid research clinics coronary primary prevention trial results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. JAMA 1984;251:369–74.
15 Holme I. An analysis of randomised trials evaluating the effect of cholesterol reduction on total mortality and coronary heart disease incidence. Circulation 1990;82:1916–24.
16 Rossouw JE, Lewis B, Rifkind BM. The value of lowering cholesterol after myocardial infarction. N Engl J Med 1990;323:1112–9.
17 McCormick J, Skrabanek P. Coronary heart disease is not preventable by population interventions. Lancet 1988;ii:839–41.
18 Muldoon MF, Manuck SB, Matthews KA. Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. Br Med J 1990;301:309–14.
19 Ravnskov U. Cholesterol-lowering trials in coronary heart disease: frequency of citation and outcome. Br Med J 1992;305:15–9.
20 Peto R. Review of primary and secondary prevention trials (lecture). Second international conference on preventive cardiology, Washington USA 1989 (quoted in reference 21).
21 Standing Medical Advisory Committee. Blood cholesterol testing: cost effectiveness of opportunistic cholesterol testing. London: Department of Health, 1990.
22 West RR. Cholesterol in prevention and health promotion: in Stephens MR (ed) Cardiovascular disease in Wales, technical document for Welsh health planning forum. Cardiff: Welsh Office, 1991.
23 West RR. Statistical overviews (meta-analyses) of interventional trials corroborate those of observational studies: example of cholesterol (lecture). Sheffield: Royal Statistical Society triennial conference, 1992.
24 West RR. Assessment of evidence versus consensus or prejudice (editorial) J Epidemiol Community Health 1992;46:321–2.
25 Jacobs D, Blackburn H, Higgins M, et al. Report on conference on low blood cholesterol: mortality associations. Circulation 1992;86:1046–60.
26 Holmes JA, West RR. Effect of P120 expression on outcome in acute myeloblastic leukaemia (submitted).

Address for correspondence: Dr R R West, Reader in Epidemiology, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN.