STATISTICS IN MEDICINE, VOL. 6, 233–240 (1987)

WHY DO WE NEED SYSTEMATIC OVERVIEWS OF RANDOMIZED TRIALS?

(Transcript of an oral presentation, modified by the editors)

RICHARD PETO

University of Oxford

My purpose in this talk is to explain why systematic overviews of randomized trials are needed and to discuss some of the problems involved in conducting such overviews.

Two medical principles that govern clinical trial design in general—and overviews in particular—are as follows. First, moderate effects on mortality may be humanly worthwhile. Take acute myocardial infarction as an example. Every year millions of patients are admitted to coronary care units with heart attacks. Some hundreds of thousands of these are going to die in hospital or within a few months of discharge. If we could reduce these deaths even quite moderately, say by 10 per cent, with a widely practicable treatment, something that could be used in the ordinary hospitals rather than just in 'high tech' hospitals, then we could save some tens of thousands of lives. However, it is horrendously difficult to detect a 10 per cent—or even a 20 per cent—reduction in the risk of death. Studies involving at least a thousand, and preferably a few thousand, *deaths* are required to detect such effects. Consider, for example, an evenly randomized trial involving 'only' one thousand deaths. If treatment has no material effect on survival, you would expect about 500 deaths in each group—give or take the effects of the play of chance. If, on the other hand, treatment really does prevent 10 per cent of all deaths, then you might expect 500 deaths in the control group and 450 in the treated group. This result would be of only marginal statistical significance. Indeed, even if you are satisfied with detecting 20 per cent risk reductions, you need at least a thousand deaths to obtain reliable answers. You do not want false negatives when you are evaluating treatments that might prevent tens of thousands of deaths a year. At the moment, the trials we do are not large enough to answer the questions we want to answer as reliably as we would want to answer them. That is one important reason for doing overviews of randomized trials.

A second principle of clinical trial design and analysis—which is the source of an enormous amount of unnecessary confusion—is that although therapeutic effects in different circumstances may very well be different, these differences are likely to be in the size of any effects rather than in their direction. If you do trials in different countries, in different populations, in different age groups or with different treatment regimens, it would be extraordinary if the risk reductions were exactly the same in every case. You would expect differences among the trials. If the studies were big enough you would be able to measure these differences reliably, but in most cases this will not be possible. The situation that treatment reduces risk in one category of patients and increases risk in another category of patients would be unusual, although certainly not impossible. Risk reductions are more likely to differ in size than in direction, and in some senses this is the fundamental assumption underlying overviews of trials. For, while we cannot assume that different

0277–6715/87/030233–08$05.00

trials are exactly comparable, or that patients in different trials are exactly comparable, it *is* reasonable to assume that if different trials address related questions then there is going to be some tendency for the answers to come out in the same direction. That tendency may well be obscured in individual trials, or even in some cases reversed, by the play of chance. But, elsewhere it may remain, and it is that tendency that the overview is trying to detect. In performing overviews, we are not trying to provide exact quantitative estimates of percentage risk reductions in some precisely defined population of patients. We are simply trying to determine whether or not some type of treatment—tested in a wide range of trials—produces any effect on mortality. If it does, then the effect seen in the trials is likely to generalize at least qualitatively to the even more loosely defined range of patients to whom the trial results are likely to be applied in the real world. Overviews provide guidance as to the approximate effect of treatment in patients studied in the trials and more importantly—to the likely effect of treatment in future patients treated outside of the trials.

If we are trying to measure moderate effects on mortality, we cannot afford moderate random errors and we cannot afford moderate biases. We have to avoid both. First, how are we going to avoid random errors? How are we going to control the play of chance sufficiently to be able to measure risk reductions of 10 to 25 per cent? Obviously, as for an individual trial, we need large numbers of patients. As we have discussed before, ideally we need studies involving a total of a few thousand deaths if we are to be able to detect moderate reductions in mortality. Few, if any, trials have recorded this many deaths. Another way to help reduce random error is to split the causes of death into those likely and unlikely to be affected by the treatment, and to do disease-specific analyses. For, if a treatment reduces the risk of death from heart disease, you cannot afford to dilute your results with fluctuations owing to deaths from other causes (e.g. traffic accidents, etc.), unless you have an infinitely large trial or an extremely large effect. Of course deaths from all causes should be reported, but inference should chiefly be based on an awareness of disease-specific analyses. Quite often, total mortality analyses may be non-significant even when the disease-specific analyses are very clearly significant and very clearly informative. This may be the case when treatment, as expected, reduces mortality only from some particular cause, but it may also be the case when treatment produces some serious—and perhaps unexpected—side-effect. As an example, consider the use of diethylstilbesterol (DES) as a treatment for prostate cancer. DES reduces the risk of death from prostate cancer but increases the risk of death from heart disease. Without disease-specific analyses, the hazard that this treatment confers on patients at high risk of death from MI could easily have been missed. In infinitely large trials, of course, it would not be necessary to worry about controlling random error in this way. However, even in a large overview, one does not have infinite statistical power. Consequently, important treatment effects can be missed if only total mortality is considered.

In addition to disease-specific analyses, one can help control random errors by basing inference chiefly on an overview of *all* patients studied in *all* related unbiased trials. By unbiased I mean that the treatment was assigned according to a random mechanism such that investigators would not have foreknowledge of the treatment allocation.

Now, let us consider avoidance of moderate biases: if you are trying to detect moderate differences, you cannot afford moderate biases. Randomization avoids the risk of bias. Unless you randomize, you cannot know that you have avoided bias and nobody else will know that you have avoided it. Randomization is not essential if you are studying a treatment that has a huge effect, but it is when you are looking for moderate differences as is the objective of overviews. Therefore, overviews do have to be restricted to the randomized trials.

In avoiding bias, it is helpful to define an objective and unbiasedly ascertained endpoint; 'death' works very well. However, bearing in mind the conflict between the need for objectivity and the point about disease-specific analyses, what we really need are categories of disease-specific

analyses that are objective, unbiasedly ascertained, and unbiasedly selected, preferably in the protocol at the outset of the trial before we come to analyse the results. Although data-dependent analyses may be useful to *generate* hypotheses for study in future trials, they should be interpreted extremely cautiously. One can always find data-derived subgroup findings—funny little interactions—in individual trial results, and in overviews. Thus, it will often appear that treatment seems to work better in one patient category than in some other category. This may be true, but because of the large probability that one (or more) of a large number of interactions looked at will be positive—whether or not there really is any treatment effect—most of the subgroup analyses from individual trials or from overviews of randomized trials should be just reported, but not believed.

To ensure the final analysis is unbiased, it is important to compare outcome among all those allocated treatment with all those allocated control (i.e. 'intention-to-treat' analyses), even if some patient did not actually receive the allocated treatment. So, outcome among all randomized patients is required, even if they are non-compliant, or if they move away—losses to follow-up must be avoided.

Finally for the control of systematic errors, inference should be based chiefly on an overview of all related unbiased trials. We cannot leave ourselves open to the bias of selecting trials for use in an overview just because we like the results. If we do not obtain the totality of the evidence, we are likely to have a biased selection of it, and—as has been pointed out—such data-derived subgroup findings are always to be viewed with suspicion.

Overviews of trials are sometimes portrayed as machines for generating positive results. This is not true: overviews of treatments that do not work produce "null" results. An example of this involves radiotherapy for women who have just had mastectomy for early breast cancer. The question is whether irradiation of the operative area and the axilla following mastectomy reduces the risk of death. An overview of all randomized trials on this subject included some 12,000 women and showed conclusively that there is no material effect of this treatment on 15-year survival. Thus, the overview provides very clear evidence that the treatment has no positive effect. Overviews are not machines for generating positive results; they are machines for generating accurate results.

At this point, I would like to discuss the method of analysing the results from many trials of a particular question in an overview. This involves the simple comparison of the number of deaths *observed* among the treatment-allocated patients in each trial with the number *expected* in that particular trial under the null hypothesis of no treatment effect. For example, consider the overview we conducted of the randomized trials of beta-blockers used long-term in the months or years following a heart attack. The two largest trials were the practolol trial (ref. 5.4 in Table I), in which the observed minus expected was -13 with variance 53 (not conventionally significant, but promising) and the timolol trial (ref. 5.7), which was conventionally significant in favour of propranolol. (Note that an observed minus expected of -13 suggests avoidance of about 26 deaths.) A dozen other rather smaller trials with a similar design were also included. Another eight trials had a slightly different design; one might want to look at them separately. For each trial we calculated observed minus expected deaths. In each case, the observed minus expected could equally well be positive or negative if treatment had no effect. When we added together all of the observed minus expected numbers, however, we got a grand total of -105, suggesting the avoidance of about 200 deaths. The basis for inference here has nothing to do with comparability of patients across trials or comparability of effect size across trials. The argument is simply that if none of these treatments had any effect, then each of the $(O - E)$s could equally well have been positive or negative, and their sum would likely have been close to zero. But, if treatment really reduced mortality then the $(O - E)$s would tend to be negative, and although this tendency might be obscured or even reversed by the play of chance in individual studies, it might stand out clearly

Table I. Total Mortality from Long-term Trials with Treatment Starting Late, and Mortality from Day 8 Onwards in Long-term Trials that Began Early and Continued After Discharge*

(Table 10 from *Progress in Cardiovascular Diseases*, **27**, 335–371 (1985), 'Beta-blockade during and after myocardial infarction: an overview of the randomized trials', by permission of Grune and Stratton Inc.)[1]

	Trial (and duration)	Basic Data from Trials (deaths/no. randomized) Allocated Beta Blocker		Allocated Control		Ratio of Percentages	Statistical Calculations: Observed minus Expected $(O-E)$	Variance of $(O-E)$	P (two-sided)
Late-Entry Trials									
5.1†	Reynolds[91] (1 y)	3/ 38	(8%)	3/ 39	(8%)	1·0	0·0	1·4	NS
5.2	Wilhelmsson[92] (2 y)	7/ 114	(6%)	14/ 116	(12%)	0·5	−3·4	4·8	NS
5.3†	Ahlmark[93] (2 y)	5/ 69	(7%)	11/ 93	(12%)	0·6	−1·8	3·5	NS
5.4	Multicentre Int.[94] + personal communication (1–3 y)	102/1,533	(7%)	127/1,520	(8%)	0·8	−13·0	53·0	0·08
5.5	Baber[95] (3–9 mo)	28/ 355	(8%)	27/ 365	(7%)	1·1	+0·9	12·7	NS
5.6	Rehnqvist[96] + personal communication (1 y)	4/ 59	(7%)	6/ 52	(12%)	0·6	−1·3	2·3	NS
5.7	Norwegian Multicentre[97] (1–3 y)	98/ 945	(10%)	152/ 939	(16%)	0·6	−27·4	54·2	0·0002
5·8	Taylor[99] (mean 4 y)	60/ 632	(9%)	48/ 471	(10%)	0·9	−1·9	23·9	NS
5·9	Hansteen[100] (1 y)	25/ 278	(9%)	37/ 282	(13%)	0·7	−5·8	13·8	NS
5.10	BHAT[101] (median 2 y)	138/1,916	(7%)	188/1,921	(10%)	0·7	−24·8	74·6	0·004
5.11	Julian[102] (1 y)	64/ 873	(7%)	52/ 583	(9%)	0·8	−5·6	25·6	NS
5.12	Australian/Swedish[103] (2 y)	45/ 263	(17%)	47/ 266	(18%)	1·0	−0·7	19·0	NS
5.13	Manger Cats[104] (1 y)	9/ 291	(3%)	16/ 293	(5%)	0·6′	−3·5	6·0	NS
5.14	EIS[105] (1 y)	57/ 858	(7%)	45/ 883	(5%)	1·3	+6·7	24·0	NS
5.15	Rehnqvist[106] (3 y)	25/ 154	(16%)	31/ 147	(21%)	0·8	−3·7	11·4	NS
5.16	Ciba-Geigy[107]	Trial closed (with 2,400 patients): data not yet available.							
	Subtotal: Late-entry trials	670/8,378	(8·0%)‡	804/7,970	(10·1%)‡	Not calculated‡	−85·1	330·3	<0.0001
Late Mortality in early-entry trials§									
2.1	Barber[56] (2 y)	33/ 207	(16%)	38/ 213	(18%)	0·9	−2·0	14·8	NS
2.2	Yusuf[60]	1/ 11	(9%)	1/ 11	(9%)	1·0	0·0	0·5	NS
2.3	Wilcox 1[57] (1 y)	28/ 251	(11%)	12/ 122	(10%)	1·1	+1·1	7·9	NS
2.4	Wilcox 2[58] (6 wk)	8/ 151	(5%)	6/ 154	(4%)	1·4	+1·1	3·4	NS
2.5	CPRG[59] (8 wk)	6/ 174	(3%)	3/ 134	(2%)	1·5	+0·9	2·2	NS
4.1	Andersen[83] (1 y)	32/ 209	(15%)	40/ 218	(18%)	0·8	−3·2	15·0	NS
4.2	Salathia[86a] (1 y)	27/ 391	(7%)	43/ 364	(12%)	0·6	−9·3	15·9	0·02
4.3	Hjalmarson[87] (13 wk)	22/ 680	(3%)	39/ 674	(6%)	0·6	−8·6	14·6	0·02

Subtotal:								
Late mortality in early-entry trials	157/ 2,074	(7·6%)‡	182/1,860	(9·8%)‡	Not calculated‡	−20·1	74·1	<0·1
Total:								
Late mortality	827/10,452	(7·9%)‡	986/9,860	(10·0%)‡	Not calculated‡	−105·2	404·5	<0·0001

Typical odds ratio of 0·77 with standard error ± 0·04 and with 95% confidence intervals for the typical odds ratio running from 0·70–0·85, which is an unusually narrow range of uncertainty ($\chi^2_1 = 27{\cdot}4$, $P < 0{\cdot}0001$). Tests for heterogeneity: the 15 purely long-term trials suggest a pooled odds ratio of 0·77, the data in this table on the eight other trials suggest a similar pooled odds ratio (0·76), and a chi-square test (on 22 df) for heterogeneity of the 23 results about this overall pooled value yields $\chi^2_{22} = 23{\cdot}3$, NS. Note that the EIS (5.14) stopped early because of adverse trends, which will increase the heterogeneity slightly and which invalidates a conventional test of whether the effects in that study were significantly worse than in the aggregate of the remaining 22 studies. Among these remaining 22, however, none has results that point to an odds ratio significantly better or significantly worse than the overall pooled odds ratio of 0·77.

* The trial of Davies[130] in which 1/35 propranolol-allocated and 2/33 placebo-allocated patients died has been omitted as a large proportion of the patients in it had not suffered an infarction. The trial of Mazur *et al.*[145] is also omitted, as despite its abstract[146] it was not randomized.

† Trials 5.1 and 5.3 suffered from extensive postrandomization withdrawals.

§ Deaths after day 7/survivors after days 0–7.

‡ In this table, direct comparison of patients in one trial with patients in another may be inappropriate, so an overview of the results of several trials should be based on the sum of the $O - E$ values from the separate trials rather than on the overall percentages, and the pooled odds ratio derives from this sum.

Table II. Actual Effects of Trial Size on Trial Results

Relationship between the total number of deaths in the two treatment groups and the result actually attained, in the 24 trials of a treatment (long-term beta blockade) that reduces the odds of death by about one-quarter.

(Table 19 from *Progress in Cardiovascular Disease*, **27**, 335–371 (1985). 'Beta blockade during and after myocardial infarction: an overview of the randomized trials', by permission of Grune and Strattan, Inc.)[1]

Total Deaths in Trial (beta blocker + placebo)	Mean No. Patients Randomized	Statistical Power	Trial Results			
			$P < 0{\cdot}05$ against	NS against	NS favorable†	$P < 0{\cdot}05$ favorable
0–50	255	Utterly inadequate	0	5	5	0
50–150	861	Probably inadequate	0	1	8	2
150–350	2,925	Possibly adequate, possibly not	0	0	1	2
350–650	No such beta blocker trials exist	Probably adequate	—	—	—	—
>650	No such beta blocker trials exist	Definitely adequate	—	—	—	—
Total	866	Inadequate separately, adequate only in aggregate	0	6	14	4

† Includes one very small trial with zero difference.

when the grand total of the $(O - E)$s from all of the trials is examined. The grand total in this overview was -100, five standard deviations away from zero. This demonstrates that the effect is clearly real.

But why use observed minus expecteds rather than some logistic model? One nice thing about $(O - E)$s is that the analysis is readily understandable to physicians who actually need to know when to use and when to avoid using beta-blockers. Logistic regression models, maximum likelihood, and Cox regression are not so easily understood. These models may have a certain mathematical appeal, but they are not necessary; an analysis based on $(O - E)$s is optimally sensitive to any real effects that may await discovery (Yusuf *et al.*, 1985). One can get all of the asymptotic efficiency of logistic regression or Cox regression by the use of crude $(O - E)$ numbers, while avoiding the assumption that the relative risk is the same in each trial, or that a proportional hazards model is appropriate. Therefore, from the statistical viewpoint, the simple sum of $(O - E)$s is not only more understandable, it is also more justifiable. It is both clearly unbiased and assumption free.

The second attractive aspect of the $(O - E)$ approach is that it leads not only to a test of the null hypothesis, but also to a description of the alternative hypothesis. Consider also the previous example in which we had an $(O - E)$ of -13 with a variance of 53. The ratio $-13/53$ is actually a remarkably good estimate of the log odds ratio. The point is that although the sum of $(O - E)$s from separate trials looks crude it can be used in a sophisticated way to provide a useful description of the alternative hypothesis. It should be stressed that in doing this, there is no assumption about homogeneity among these trials; one is simply calculating what we like to call the 'typical' odds ratio. To clarify this concept, suppose we measured a haemoglobin level for everybody attending this conference and calculated the 'sex effect', i.e. the difference between the male and the female average. In making this calculation there would be no assumption that we all have the same haemoglobin level. We would merely be getting a 'typical' sex effect value for scientists interested in the methodology of overviews. Of course, this value is likely to be generalizable beyond the narrow population in which haemoglobin has been measured to other populations—such as scientists that are not interested in overviews, or to non-scientists. Similarly, in overviews of trials, the 'typical' odds ratio provides a useful approximation to the likely size of any effect of treatment in patients other than these studied in the trials. The degree to which such results can reasonably be extrapolated, however, must remain a matter of judgement rather than one of precise mathematical calculation.

Odds ratios can be used to describe the trial results pictorially, as in Figure 1. The vertical solid line represents an odds ratio of 1·0, i.e. no treatment effect. Each horizontal line represents the point estimate and 95% confidence interval for each separate trial (or for an overview of the very small trials). The timolol trial (5·7) looks very promising. The oxprenolol trial (5·14) looks very unpromising. The vertical dotted line represents the odds ratio suggested by the overview of all of those trials. This is an odds ratio of 0·8—suggesting a reduction in the odds of death of about 20 per cent. Notice, though, the substantial uncertainty of the individual trial results; these trials are trying to measure a 20 per cent risk reduction with standard errors that are typically bigger than 20 per cent. Not surprisingly, of the 24 trials in the published overview, 20 yielded differences that are not conventionally significant (Table II). That is not chiefly because of heterogeneity (there is no statistically significant heterogeneity among them, although one would not really believe that the trials were entirely homogeneous) but because the probability of getting a false negative, if you try and measure this kind of difference with a conventionally sized trial, will be large.

I would now like to say a word about interactions. When we reviewed these trials of long-term beta-blockade, we tried to see if some categories of beta-blockers were more effective than others. First we split them as to whether they were cardioselective or not; that did not seem to make any

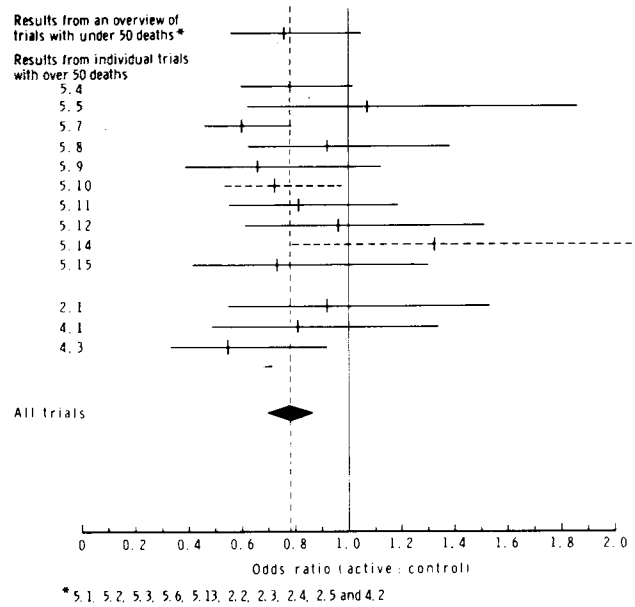

Figure 2. Mortality in long-term beta blocker trials, by ancillary properties of agent tested: odds ratios (active:control), estimated as in Table 13, together with approxinate 95% confidence ranges.

(Figure 6 from *Progress in Cardiovascular Diseases*, **27**, 335–371 (1985). 'Beta blockage during and after myocardial infarction: an overview of the randomized trials' by permission of Gune and Stratton, Inc.)[1]

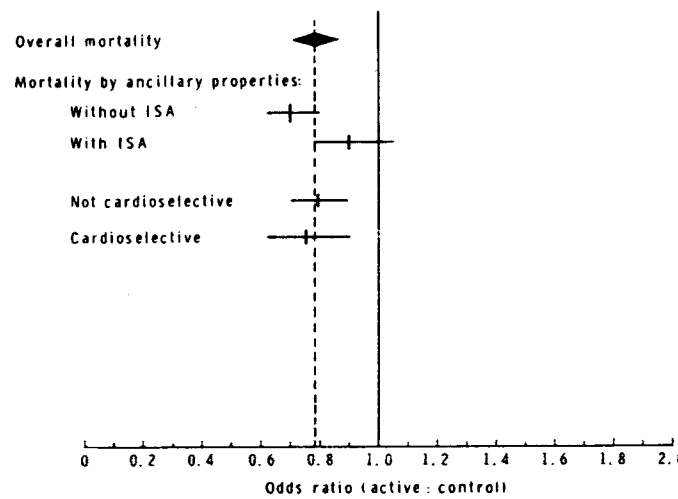

Figure 1. Mortality by allocated treatment in all the randomized trials of long-term beta blockade following myocardial infarction: odds ratios (active:control), together with approximately 95 per cent or 99 per cent confidence ranges. —|— 95 per cent confidence range for trials that ran to scheduled finish, – – – – – 99 per cent confidence range for trials stopped early due to good/bad trend, ◆ 95 per cent confidence range from an overview of all the trials.

(Figure 4 from *Progress in Cardiovascular Diseases*, **27**, 335–371 (1985), 'Beta blockade during and after myocardial infarction: an overview of the randomized trials', by permission of Grune and Stratton, Inc.)[1]

difference. Then we split them as to whether they had intrinsic sympathomimetic activity (ISA), and lo and behold, we found that the agents without ISA seemed to produce a bigger effect than the agents with ISA (Figure 2). This difference was conventionally significant at the 0·01 level. At the time this seemed rather impressive—and it did not take long to think up a biological 'explanation' for it—but, it is interesting that all the data that has turned up since has tended to contradict this finding. When we looked at the non-fatal reinfarctions, there was no difference at all between the agents with ISA and the agents without ISA; overall, they both produced the same 20 per cent reduction. More recently we have seen results of two more trials, one of an agent with ISA and one of an agent without ISA. These trials showed the opposite of our initial observation: the agent without ISA had one of the most unpromising results ever seen, and the agent with ISA had one of the most promising results ever seen. These two extra trials demolish that statistically significant interaction. In retrospect, we were wrong to give it as much credence as we did. It was right to observe it and report it; but it was wrong to believe it. This is a terribly important point and is a recurring theme in the analysis of individual trials and of overviews.

To underline the necessity of performing disease-specific analyses in order to reduce random error in overviews, I would like to consider an overview of trials of blood pressure reduction. There were 486 strokes in the collective control groups and 289 in the collective treated groups. That is a difference of 6·6 standard deviations, a massively significant 40 per cent reduction. There was also a moderate effect on heart attacks (785 versus 697, about 3 standard deviations). Because most vascular deaths involve either heart disease or stroke, we can combine these end points for a 3·3 standard deviation difference in total vascular mortality. The total number of non-vascular deaths (that is cancers, traffic accidents, and so on) was 400 versus 389, minutely favourable but not statistically significant. Of course, when everything is added up, the total mortality is also statistically significantly reduced, but it is interesting to note that if the non-vascular mortality had come out differently—if the trial results had just happened to be one standard deviation against treatment in non-vascular mortality—total mortality would not have been significantly improved. Nevertheless, the interpretation of the overview should be the same. This is a very important point; basing inference on total mortality is not really reliable. Given an infinite amount of data, the total mortality analysis would be very interesting, but with a limited amount of data the disease specific analyses are likely to be more informative. Not only are they more sensitive to any real effects, but they are also more informatively generalizable to different populations.

In summary, one needs the overview for the large picture. In overviews, as in individual trials, the minute details—the small interactions, the moderately significant subset effects—really have to be distrusted. It is the large picture that is important. In my view, that is best achieved from overviews of all randomized trials of a particular question, basing inference principally on disease specific analyses, but with data on total mortality also clearly available.

REFERENCES

1. Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, T. 'Beta blockade during and after myocardial infarction: and overview of the randomized trials', *Progress in Cardiovascular Disease*, **27**, 335–371 (1985).
2. Yusuf, S., Collins, R. and Peto, R. 'Why do we need some large, simple randomized trials?', *Statistics in Medicine*, **3**, 409–420 (1984).