Meta-analysis
Spurious precision? Meta-analysis of observational studies

Matthias Egger, Martin Schneider, George Davey Smith

This is the fifth in a series of six articles examining the procedures in conducting reliable meta-analysis in medical research

Department of Social Medicine, University of Bristol, Bristol BS8 2PR
Matthias Egger, *reader in social medicine and epidemiology*
George Davey Smith, *professor of clinical epidemiology*

Department of Social and Preventive Medicine, University of Berne, CH-3012 Berne, Switzerland
Martin Schneider, *research fellow*

Correspondence to: Dr Egger m.egger@bristol. ac.uk

BMJ 1998;316:140–4

In previous articles we have focused on the potentials, principles, and pitfalls of meta-analysis of randomised controlled trials.[1-5] Meta-analysis of observational data is, however, also becoming common. In a Medline search we identified 566 articles (excluding those published as letters) published in 1995 and indexed with the medical subject heading (MeSH) term "meta-analysis." We randomly selected 100 of these articles and examined them further. Sixty articles reported on actual meta-analyses, and 40 were methodological papers, editorials, and traditional reviews (table). Among the meta-analyses, about half were based on observational studies, mainly cohort and case-control studies of medical interventions or aetiological associations.

Characteristics of 100 articles randomly selected from articles published in 1995 and indexed in Medline with keyword "meta-analysis"

Type of article	No of articles
Meta-analysis of controlled trials	33
Meta-analysis of observational studies*	27
Methodological article	21
Editorial or commentary	9
Traditional review	6
Other	4

Letters were excluded.
*12 meta-analyses of studies of aetiological associations, 11 of therapeutic interventions, and 4 of diagnostic procedures.

The randomised controlled trial is the principal research design in the evaluation of medical interventions. However, aetiological hypotheses—for example, those relating common exposures to the occurrence of disease—cannot generally be tested in randomised experiments. Does breathing other people's tobacco smoke cause lung cancer, drinking coffee cause coronary heart disease, and eating a diet rich in saturated fat cause breast cancer? Studies of such "menaces of daily life"[6] use observational designs or examine the presumed biological mechanisms in the laboratory. In these situations the risks involved are generally small, but once a large proportion of the population is exposed, the potential public health implications of these associations—if they are causal—can be striking.

Analyses of observational data also have a role in medical effectiveness research.[7] The evidence available from clinical trials will rarely answer all the important questions. Most trials are conducted to establish efficacy and safety of a single agent in a specific clinical situation. Owing to the limited size of such trials, less common adverse effects of drugs may only be detected in case-control studies or in analyses of databases from postmarketing surveillance schemes. Also, because follow up is generally limited, adverse effects occurring many years later will not be identified. If years later

Summary points

Meta-analysis of observational studies is as common as meta-analysis of controlled trials

Confounding and selection bias often distort the findings from observational studies

There is a danger that meta-analyses of observational data produce very precise but equally spurious results

The statistical combination of data should therefore not be a prominent component of reviews of observational studies

More is gained by carefully examining possible sources of heterogeneity between the results from observational studies

Reviews of any type of research and data should use a systematic approach, which is documented in a materials and methods section

established interventions are incriminated with adverse effects, there will be ethical, political, and legal obstacles to the conduct of a new trial. Recent examples for such situations include the controversy surrounding a possible association between intramuscular administration of vitamin K to newborns and the risk of childhood cancer[8] and whether oral contraceptives increase women's risk of breast cancer.[9]

The patients who are enrolled in randomised trials often differ from the average patient seen in clinical practice. Women, elderly people, and minority ethnic groups are often excluded from randomised trials.[10 11] Similarly, the university hospitals typically participating in clinical trials differ from the settings where most patients are treated. In the absence of evidence from randomised trials from these settings and patient groups, the results from observational database analyses may seem more relevant and more readily applicable to clinical practice.[12] Finally, strong prior views may preclude the recruitment of sufficient patients or clinics into a randomised experiment. In complementary medicine, for example, consider a treatment that entails drinking your own urine.[13] It would probably be impossible to recruit sufficient patients into a controlled trial.

Meta-analysis, by promising a precise and definite answer when the magnitude of the underlying risks are small or when the results from individual studies disagree, seems an attractive proposition both in aetiological studies and in observational effectiveness research.
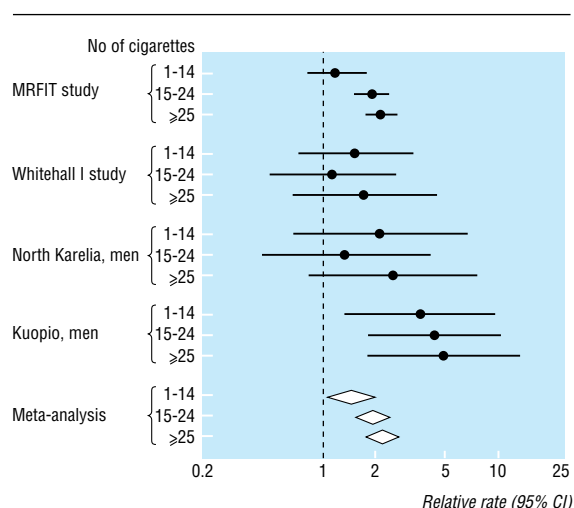
Fig 1 Adjusted relative rates of suicide among middle aged male smokers compared with non-smokers. Results from four cohort studies adjusted for age plus income, race, cardiovascular disease, diabetes (multiple risk factors intervention trial (MRFIT)), employment grade (Whitehall I study), alcohol use, serum cholesterol concentration, systolic blood pressure, education (North Karelia and Kuopio studies). Meta-analysis is by fixed effects model

Critical difference in assumptions

Meta-analysis of randomised trials is based on the assumption that each trial provides an unbiased estimate of the effect of an experimental treatment, with the variability of the results between the studies being attributed to random variation. The overall effect calculated from a group of sensibly combined and representative randomised trials will provide an essentially unbiased estimate of the treatment effect, with an increase in the precision of this estimate. A fundamentally different situation arises in the case of observational studies. Such studies yield estimates of association which may deviate from true underlying relationships beyond the play of chance. This may be due to the effects of confounding factors, the influence of biases, or both.

Confounding, residual confounding, and bias

Patients exposed to the factor under investigation may differ in several other aspects that are relevant to the risk of developing the disease in question. Consider, for example, smoking as a risk factor for suicide. Virtually all cohort studies have shown a positive association, with a dose-response relation being evident between the amount smoked and the probability of committing suicide.[14-19] Figure 1 illustrates this for four prospective studies of middle aged men, including the massive cohort of patients screened for the multiple risk factors intervention trial. Based on over 390 000 men and almost five million years of follow up, a meta-analysis of these cohorts produces highly precise and significant estimates of the increase in suicide risk that is associated with smoking different daily amounts of cigarettes: relative rate for 1-14 cigarettes 1.43 (95% confidence interval 1.06 to 1.93), for 15-24 cigarettes 1.88 (1.53 to 2.32), ≥25 cigarettes 2.18 (1.82 to 2.61).

On the basis of established criteria,[20] many would consider the association to be causal—if only it were more plausible. Indeed, it is improbable that smoking is causally related to suicide.[14] Rather, it is the social and mental states predisposing to suicide that are also associated with the habit of smoking. Factors that are related to both the exposure and the disease under study—confounding factors—may thus distort results. If the factor is known and has been measured, the usual approach is to adjust for its influence in the analysis. For example, studies assessing the influence of coffee consumption on the risk of myocardial infarction should make statistical adjustments for smoking, as smoking is generally associated with drinking larger amounts of coffee, and smoking is a cause of coronary heart disease.[21] However, even if adjustments for confounding factors have been made in the analysis, residual confounding remains a potentially serious problem in observational research. Residual confounding arises when a confounding factor cannot be measured with sufficient precision—which often occurs in epidemiological studies.[22 23] Confounding is the most important threat to the validity of results from cohort studies, whereas many more difficulties, in particular selection biases, arise in case-control studies.[24]

Plausible but equally spurious findings

Implausibility of results, as in the case of smoking and suicide, rarely protects us from reaching misleading claims. It is generally easy to produce plausible explanations for the findings from observational research. In a cohort study of sex workers, for example, one group of researchers that investigated cofactors in transmission of HIV among heterosexual men and women found a strong association between oral contraceptives and HIV infection, which was independent of other factors.[25] The authors hypothesised that, among other mechanisms, the risk of transmission could be increased with oral contraceptives due to "effects on the genital mucosa, such as increasing the area of ectopy and the potential for mucosal disruption during intercourse." In a cross sectional study another group produced diametrically opposed findings, indicating that oral contraceptives protect against the
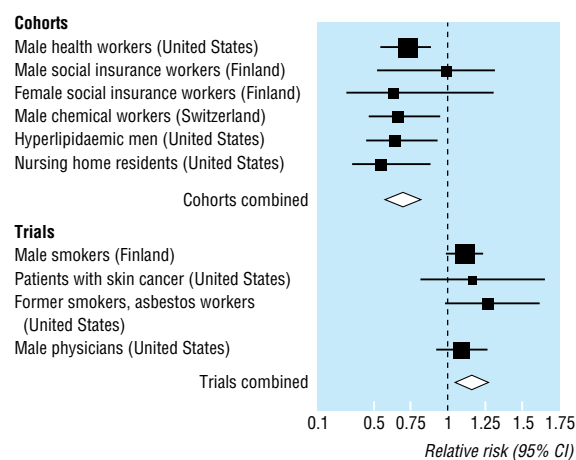

Fig 2 Meta-analysis of association between β carotene intake and cardiovascular mortality: results from observational studies show considerable benefit, whereas the findings from randomised controlled trials show an increase in the risk of death. Meta-analysis is by fixed effects model

virus.[26] This was considered to be equally plausible, "since progesterone-containing oral contraceptives thicken cervical mucus, which might be expected to hamper the entry of HIV into the uterine cavity." It is likely that confounding and bias had a role in producing these contradictory findings. This example should be kept in mind when assessing other seemingly plausible epidemiological associations.

Rare insight: protective effect of β carotene that wasn't

Observational studies have consistently shown that people eating more fruits and vegetables, which are rich in β carotene, and people having higher serum β carotene concentrations have lower rates of cardiovascular disease and cancer.[27] β carotene has antioxidant properties and could thus plausibly be expected to prevent carcinogenesis and atherogenesis by reducing oxidative damage to DNA and lipoproteins.[27] Contrary to many other associations found in observational studies, this hypothesis could be, and was, tested in experimental studies. The findings of four large trials have recently been published.[28-31] The results were disappointing and even—for the two trials conducted in men at high risk (smokers and workers exposed to asbestos)[28 29]—disturbing.

We performed a meta-analysis of the findings for cardiovascular mortality, comparing the results from the six observational studies recently reviewed by Jha et al[27] with those from the four randomised trials. For the observational studies the results relate to a comparison between groups with high and low β carotene intake or serum β carotene concentration, whereas in the trials the participants randomised to β carotene supplements were compared with those randomised to placebo. With a fixed effects model, the meta-analysis of the cohort studies shows a significantly lower risk of cardiovascular death (relative risk reduction 31% (95% confidence interval 41% to 20%, P < 0.0001)) (fig 2). The results from the randomised trials, however, show a moderate adverse effect of β carotene supplementation (relative increase in the risk of cardiovascular death 12% (4% to 22%, P = 0.005)). Similarly discrepant results between epidemiological studies and trials were observed for the incidence of and mortality from cancer. This example illustrates that in meta-analyses of observational studies, the analyst may well be simply producing tight confidence intervals around spurious results.

Case studies: exploring sources of heterogeneity

Some observers suggest that meta-analysis of observational studies should be abandoned altogether.[32] We disagree, but we think that the statistical combination of studies should not generally be a prominent component of reviews of observational studies. The thorough consideration of possible sources of heterogeneity between observational study results will provide more insights than the mechanistic calculation of an overall measure of effect, which will often be biased.

Heterogeneity can be explored in funnel plots, a graphical method discussed in detail previously.[5] Funnel plots will, however, generally be less useful in the context of observational meta-analyses. Publication bias and related biases[4] will be less important against the background of the numerous other biases and confounding factors that may introduce heterogeneity. Several such situations are depicted in figure 3. Consider diet and breast cancer. The hypothesis from ecological analyses[33] that higher intake of saturated fat could increase the risk of breast cancer generated much observational research, often with contradictory results. A comprehensive meta-analysis[34] showed an association for case-control but not for cohort studies (odds ratio 1.36 for case-control studies versus relative rate 0.95 for cohort studies comparing highest with lowest category of saturated fat intake, P = 0.0002 for difference in our calculation) (fig 2). This discrepancy was also shown in two separate large collaborative pooled analyses of cohort and case-control studies.[35 36] The most likely explanation for this situation is that biases in the recall of dietary items and in the selection of study participants have produced a spurious association in the case-control comparisons.[36]

That differential recall of past exposures may introduce bias is also evident from a meta-analysis of case-control studies of intermittent sunlight exposure and melanoma (fig 3).[37] When studies were combined in which some degree of blinding to the study hypothesis was achieved, only a small and non-significant effect (odds ratio 1.17 (95% confidence interval 0.98 to
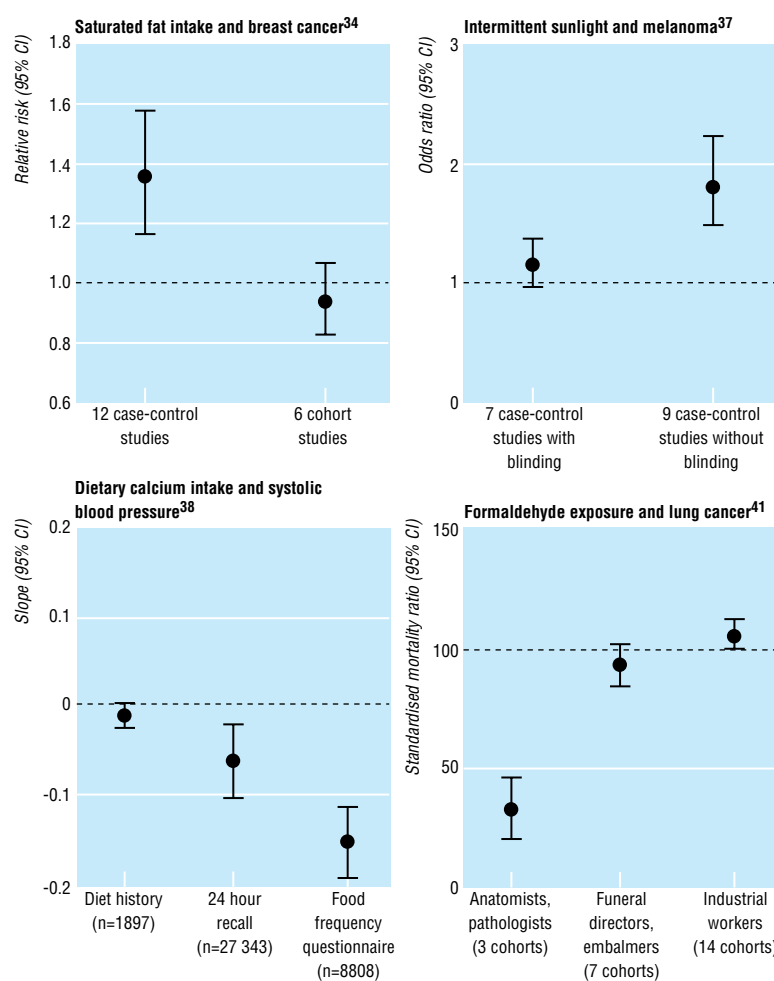

Fig 3 Examples of heterogeneity in published observational meta-analyses

MARK HUDSON

1.39)) was evident. Conversely, in studies without blinding, the effect was considerably greater and significant (1.84 (1.52 to 2.25)). The difference between these two estimates is unlikely to be a product of chance (P = 0.0004 in our calculation).

The importance of the methods used for assessing exposure is further illustrated by a meta-analysis of cross sectional data of dietary calcium intake and blood pressure from 23 different studies.[38] As shown in figure 3, the regression slope describing the change in systolic blood pressure (in mm Hg) per 100 mg of calcium intake is strongly influenced by the approach used for assessing the amount of calcium consumed. The association is small and only marginally significant with diet histories (slope − 0.01 (− 0.003 to − 0.016)) but large and highly significant when food frequency questionnaires were used (− 0.15 (− 0.11 to − 0.19). With studies using 24 hour recall an intermediate result emerges (− 0.06 (− 0.09 to − 0.03). Diet histories assess patterns of usual intake over long periods of time and require an extensive interview with a nutritionist, whereas 24 hour recall and food frequency questionnaires are simpler methods that reflect current consumption.[39] It is conceivable that different precision in the assessment of current calcium intake may explain the differences in the strength of the associations found, a statistical phenomenon known as regression dilution bias.[40]

An important criterion supporting causality of associations is a dose-response relation. In occupational epidemiology the quest to show such an association can lead to very different groups of employees being compared. In a meta-analysis that examined the link between exposure to formaldehyde and cancer, funeral directors and embalmers (high exposure) were compared with anatomists and pathologists (intermediate to high exposure) and with industrial workers (low to high exposure, depending on job assignment).[41] As shown in figure 3, there is a striking deficit of deaths from lung cancer among anatomists and pathologists (standardised mortality ratio 33 (95% confidence interval 22 to 47)), which is most likely to be due to a lower prevalence of smoking among this group. In this situation few would argue that formaldehyde protects against lung cancer. In other instances, however, such selection bias may be less obvious.

In these examples heterogeneity was explored in the spirit of sensitivity analysis[3]—to test the stability of findings across different study designs and different approaches to both exposure ascertainment and selection of study participants. Such sensitivity analyses should alert investigators to inconsistencies and prevent misleading conclusions. Although heterogeneity was noticed, explored, and sometimes extensively discussed, the way the situation was interpreted differed considerably. In the analysis examining studies of dietary fat and risk of breast cancer, the authors went on to combine case-control and cohort studies and concluded that "higher intake of dietary fat is associated with an increased risk of breast cancer."[34] The meta-analysis of exposure to sunlight and risk of melanoma was exceptional in its thorough examination of possible reasons for heterogeneity, and the calculation of a combined estimate was deemed appropriate in one subgroup of population based studies only.[37] Conversely, uninformative and potentially misleading combined estimates were calculated both in the study on dietary calcium and blood pressure[38] and in the meta-analysis of occupational exposure to formaldehyde.[41] These case studies show that the temptation to combine the results of studies seems to be hard to resist.

Conclusion

The suggestion that formal meta-analysis of observational studies can be misleading and that insufficient attention is often given to heterogeneity does not mean that researchers should return to writing highly subjective narrative reviews. Many of the principles of systematic reviews remain: a study protocol should be written in advance, complete literature searches carried out, and studies selected and data extracted in a reproducible and objective fashion.[42] This allows both differences and similarities of the results found in different settings to be inspected, hypotheses to be formulated, and the need for future studies, including randomised controlled trials, to be defined.

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* 1997;315:61-66.
5 Egger M, Davey Smith G, Schneider M, Minder CE. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.
6 Feinstein AR. Scientific standards in epidemiological studies of the menace of daily life. *Science* 1988;242:1257-63.
7 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.
8 Brousson MA, Klein MC. Controversies surrounding the administration of vitamin K to newborns: a review. *Can Med Assoc J* 1996;154:307-15.
9 Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet* 1996;347:1713-27.

10 Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA* 1992;268:1417-22.

11 Levey BA. Bridging the gender gap in research. *Clin Pharmacol Ther* 1991;50:641-6.

12 Hlatky MA. Using databases to evaluate therapy. *Stat Med* 1991;10:647-52.

13 Wilson CWM. The protective effect of auto-immune buccal urine therapy (AIBUT) against the Raynaud phenomenon. *Med Hypotheses* 1984;13:99-107.

14 Davey Smith G, Phillips AN, Neaton JD. Smoking as "independent" risk factor for suicide: illustration of an artifact from observational epidemiology. *Lancet* 1992;340:709-11.

15 Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years' observation on male British doctors. *BMJ* 1994;309:901-11.

16 Doll R, Gray R, Hafner B, Peto R. Mortality in relation to smoking: 22 years' observations on female British doctors. *BMJ* 1980;280:967-71.

17 Tverdal A, Thelle D, Stensvold I, Leren P, Bjartveit K. Mortality in relation to smoking history: 13 years follow-up of 68 000 Norwegian men and women 35-49 years. *J Clin Epidemiol* 1993;46:475-87.

18 Vartiainen E, Puska P, Pekkanen J, Tuomilehto J, Lönnqvist J, Ehnholm C. Serum cholesterol concentrations and mortality from accidents, suicide, and other violent causes. *BMJ* 1994;309:445-7.

19 Hemenway D, Solnick SJ, Colditz GA. Smoking and suicide among nurses. *Am J Pub Health* 1993;83:249-51.

20 Bradford Hill A. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295-300.

21 Leviton A, Pagano M, Allred EN, El Lozy M. Why those who drink the most coffee appear to be at increased risk of disease: a modest proposal. *Ecol Food Nutr* 1994;31:285-93.

22 Phillips AN, Davey Smith G. How independent are "independent" effects? Relative risk estimation when correlated exposures are measured imprecisely. *J Clin Epidemiol* 1991;44:1223-31.

23 Davey Smith G, Phillips AN. Confounding in epidemiological studies: why "independent" effects may not be all they seem. *BMJ* 1992;305:757-9.

24 Sackett DL. Bias in analytical research. *J Chron Dis* 1979;32:51-63.

25 Plummer FA, Simonsen JN, Cameron DW, Ndinya-Achola JO, Kreiss JK, Gakinya MN, et al. Cofactors in male-female sexual transmission of human immunodeficiency virus type 1. *J Infect Dis* 1991;233:233-9.

26 Lazzarin A, Saracco A, Musicco M, Nicolosi A. Man-to-woman sexual transmission of the human immunodeficiency virus. *Arch Intern Med* 1991;151:2411-6.

27 Jha P, Flather M, Lonn E, Farkouh M, Yusuf S. The antioxidant vitamins and cardiovascular disease. *Ann Intern Med* 1995;123:860-72.

28 Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New Engl J Med* 1994;330:1029-35.

29 Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *New Engl J Med* 1996;334:1150-5.

30 Hennekens CH, Buring JE, Manson JC, Stampfer M, Rosner B, Cook NR, et al. Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *New Engl J Med* 1996;334:1145-9.

31 Greenberg ER, Baron JA, Karagas MR, Stukel TA, Nierenberg DW, Stevens MM, et al. Mortality associated with low plasma concentration of beta carotene and the effect of oral supplementation. *JAMA* 1996;275:699-703.

32 Shapiro S. Meta-analysis/Shmeta-analysis. *Am J Epidemiol* 1994;140:771-8.

33 Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries with special reference to dietary practices. *Int J Cancer* 1975;15:617-31.

34 Boyd NF, Martin LJ, Noffel M, Lockwood GA, Tritchler DL. A meta-analysis of studies of dietary fat and breast cancer. *Br J Cancer* 1993;68:627-36.

35 Howe GR, Hirohata T, Hislop TG, Iscovich JM, Yuan JM, Katsouyanni K, et al. Dietary factors and risk of breast cancer: combined analysis of 12 case-control studies. *J Natl Cancer Inst* 1990;82:561-569.

36 Hunter DJ, Spiegelman D, Adami H-O, Beeson L, van den Brandt PA, Folsom AR, et al. Cohort studies of fat intake and the risk of breast cancer - a pooled analysis. *New Engl J Med* 1996;334:356-61.

37 Nelemans PJ, Rampen FHJ, Ruiter DJ, Verbeek ALM. An addition to the controversy on sunlight exposure and melanoma risk: a meta-analytical approach. *J Clin Epidemiol* 1995;48:1331-42.

38 Cappuccio FP, Elliott P, Allender PS, Pryer J, Follman DA, Cutler JA. Epidemiologic association between dietary calcium intake and blood pressure: a meta-analysis of published data. *Am J Epidemiol* 1995;142:935-45.

39 Block G. A review of validations of dietary assessment methods. *Am J Epidemiol* 1982;115:492-505.

40 MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, et al. Blood pressure, stroke and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990;335:765-74.

41 Blair A, Saracci R, Stewart PA, Hayes RB, Shy C. Epidemiologic evidence on the relationship between formaldehyde exposure and cancer. *Scand J Work Environ Health* 1990;16:381-93.

42 Chalmers I, Altman DG. Foreword. In: Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publishing, 1995.

A patient that changed my practice

Mother knows best

I have lost count of the number of times my teachers taught that in paediatrics "Mother is always right."

As a casualty officer I decided to do locum work. The first card I picked up was that of a 4 month old boy who had had "one week history of cough, crying all the time." Squeezed in at the end of the summary was, "Mother concerned about development." She told me that she had three other children. "I know that there is something wrong here. It's more than just that he's small and will catch up later," she added.

The boy had been born after a complication free pre-, intra-, and postpartum pregnancy and was "having regular check ups" as the mother had done a circuit of all the near and not so near accident and emergency departments. The child was asleep, had a temperature of 37.7°C, was normotensive, and did not have a tachycardia.

Respiratory, cardiovascular, abdominal, otolaryngeal, and neurological examinations were all normal, apart from a macular, blanching non-pruritic rash on his trunk. I explained that the rash was connected to the cough, would go when the cough went, and that that was why he was a little floppy. The mother insisted that none of her other three children had ever been as floppy even when asleep or ill.

It was at that stage that another of my teachers' pronouncements came at me like a bolt of lightening: "If you don't remember anything else always remember to ...." I went up to the sister and asked for "one of those graphs that paediatricians use to plot weight and head circumference." If you bear in mind that I was a locum seeing my first patient you can imagine the look I got from the regular doctors. Eventually a pristine pack of the charts was found. The child's weight was just below the second centile, but the head circumference was below the third centile. I measured and replotted; the mother twigged. "I knew his head should be bigger" and then she said, "My other son has this skin condition that they had to get a really old retired professor to sort out. When he said that this was only the second time he had ever seen this condition in 70 years I was really worried."

That was when the penny dropped. Tactfully worded questions revealed that the mother had married her first cousin. The patient and his brother were, therefore, likely to have a sex linked chromosomal abnormality and the fact that her two daughters "were spared" was scientifically explicable as well as "by God's grace."

This lady changed my practice in that now I always weigh infants and measure their skull circumference just as my teachers told me to do. For the curious the dermatological condition was *lipid proteinosis*.

Rahul Kacker, *general practice trainee, Sheffield*

We welcome filler articles up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake,* or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.