Meta-analysis
Principles and procedures

Matthias Egger, George Davey Smith, Andrew N Phillips

Meta-analysis is a statistical procedure that integrates the results of several independent studies considered to be "combinable."[1] Well conducted meta-analyses allow a more objective appraisal of the evidence than traditional narrative reviews, provide a more precise estimate of a treatment effect, and may explain heterogeneity between the results of individual studies.[2] Ill conducted meta-analyses, on the other hand, may be biased owing to exclusion of relevant studies or inclusion of inadequate studies.[3] Misleading analyses can generally be avoided if a few basic principles are observed. In this article we discuss these principles, along with the practical steps in performing meta-analysis.

Summary points

Meta-analysis should be as carefully planned as any other research project, with a detailed written protocol being prepared in advance

The a priori definition of eligibility criteria for studies to be included and a comprehensive search for such studies are central to high quality meta-analysis

The graphical display of results from individual studies on a common scale is an important intermediate step, which allows a visual examination of the degree of heterogeneity between studies

Different statistical methods exist for combining the data, but there is no single "correct" method

A thorough sensitivity analysis is essential to assess the robustness of combined estimates to different assumptions and inclusion criteria

This is the second in a series of seven 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 Primary Care and Population Sciences, Royal Free Hospital School of Medicine, London NW3 2PF
Andrew N Phillips, *professor of epidemiology and biostatistics*
Correspondence to: Dr Egger m.egger@bristol.ac.uk

BMJ 1997;315:1533–7

Observational study of evidence

Meta-analysis should be viewed as an observational study of the evidence. The steps involved are similar to any other research undertaking: formulation of the problem to be addressed, collection and analysis of the data, and reporting of the results. Researchers should write in advance a detailed research protocol that clearly states the objectives, the hypotheses to be tested, the subgroups of interest, and the proposed methods and criteria for identifying and selecting relevant studies and extracting and analysing information.

As with criteria for including and excluding patients in clinical studies, eligibility criteria have to be defined for the data to be included. Criteria relate to the quality of trials and to the combinability of treatments, patients, outcomes, and lengths of follow up. Quality and design features of a study can influence the results.[4,5] Ideally, researchers should consider including only controlled trials with proper randomisation of patients that report on all initially included patients according to the intention to treat principle and with an objective, preferably blinded, outcome assessment.[6] Assessing the quality of a study can be a subjective process, however, especially since the information reported is often inadequate for this purpose.[7] It is therefore preferable to define only basic inclusion criteria and to perform a thorough sensitivity analysis (see below).

The strategy for identifying the relevant studies should be clearly delineated. In particular, it has to be decided whether the search will be extended to include unpublished studies, as their results may systematically differ from published trials. As will be discussed in later articles, a meta-analysis that is restricted to published evidence may produce distorted results owing to such publication bias. For locating published studies, electronic databases are useful,[8] but, used alone, they may miss a substantial proportion of relevant studies.[9,10] In an attempt to identify all published controlled trials, the Cochrane Collaboration has embarked on an extensive manual search of medical journals published in English and many other languages.[11] The Cochrane Controlled Trials Register[12] is probably the best single electronic source of trials; however, citation indices and the bibliographies of review articles, monographs, and the located studies should also be scrutinised.

A standardised record form is needed for data collection. It is useful if two independent observers extract the data, to avoid errors. At this stage the quality of the studies may be rated, with one of several specially designed scales.[13,14] Blinding observers to the names of the authors and their institutions, the names of the journals, sources of funding, and acknowledgments leads to more consistent scores.[14] This entails either photocopying papers, removing the title page, and concealing journal identifications and other characteristics with a black marker, or scanning the text of papers into a computer and preparing standardised formats.[15,16]

Standardised outcome measure

Individual results have to be expressed in a standardised format to allow for comparison between studies. If the end point is continuous—for example, blood pressure—the mean difference between the treatment and control groups is used. The size of a difference, however, is influenced by the underlying population value. An antihypertensive drug, for example, is likely to have a greater absolute effect on blood pressure in overtly hypertensive patients than in borderline hypertensive patients. Differences are therefore often presented in units of standard deviation. If the end point is binary—for example, disease versus no disease, or dead versus alive) then odds ratios or

relative risks are often calculated (box). The odds ratio has convenient mathematical properties, which allow for ease in combining data and testing the overall effect for significance. Absolute measures, such as the absolute risk reduction or the number of patients needed to be treated to prevent one event,[17] are more helpful when applying results in clinical practice (see below).

Statistical methods for calculating overall effect

The last step consists in calculating the overall effect by combining the data. A simple arithmetic average of the results from all the trials would give misleading results. The results from small studies are more subject to the play of chance and should therefore be given less weight. Methods used for meta-analysis use a weighted average of the results, in which the larger trials have more influence than the smaller ones. The statistical techniques to do this can be broadly classified into two models,[18] the difference consisting in the way the variability of the results between the studies is treated. The "fixed effects" model considers, often unreasonably, that this variability is exclusively due to random variation.[19] Therefore, if all the studies were infinitely large they would give identical results. The "random effects" model[20] assumes a different underlying effect for each study and takes this into consideration as an additional source of variation, which leads to somewhat wider confidence intervals than the fixed effects model. Effects are assumed to be randomly distributed, and the central point of this distribution is the focus of the combined effect estimate. Although neither of two models can be said to be "correct," a substantial difference in the combined effect calculated by the fixed and random effects models will be seen only if studies are markedly heterogeneous.[18]

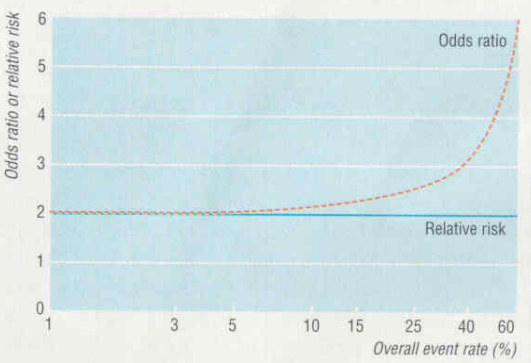
Odds ratio or relative risk?

Odds and odds ratio

The odds is the number of patients who fulfil the criteria for a given endpoint divided by the number of patients who do not. For example, the odds of diarrhoea during treatment with an antibiotic in a group of 10 patients may be 4 to 6 (4 with diarrhoea divided by 6 without, 0.66); in a control group the odds may be 1 to 9 (0.11) (a bookmaker would refer to this as 9 to 1). The odds ratio of treatment to control group would be 6 (0.66÷0.11).

Risk and relative risk

The risk is the number of patients who fulfil the criteria for a given end point divided by the total number of patients. In the example above the risks would be 4 in 10 in the treatment group and 1 in 10 in the control group, giving a risk ratio, or relative risk, of 4 (0.4÷0.1).



The odds will be close to the relative risk if the end point occurs relatively infrequently, say in less than 20%. If the outcome is more common (as in the diarrhoea example) then the odds ratio will considerably overestimate the relative risk

Bayesian meta-analysis

Some statisticians feel that other statistical approaches are more appropriate than either of the above. One approach uses Bayes's theorem, named after an 18th century English clergyman.[21] Bayesian statisticians express their belief about the size of an effect by specifying some prior probability distribution before seeing the data, and then they update that belief by deriving a posterior probability distribution, taking the data into account.[22] Bayesian models are available under both the fixed and random effects assumption.[23] The confidence interval (or more correctly in bayesian terminology, the 95% credible interval, which covers 95% of the posterior probability distribution) will often be wider than that derived from using the conventional models because another component of variability, the prior distribution, is introduced. Bayesian approaches are controversial because the definition of prior probability will often be based on subjective assessments and opinion.

Heterogeneity between study results

If the results of the studies differ greatly then it may not be appropriate to combine the results. How to ascertain whether it is appropriate, however, is unclear. One approach is to examine statistically the degree of similarity in the studies' outcomes—in other words, to

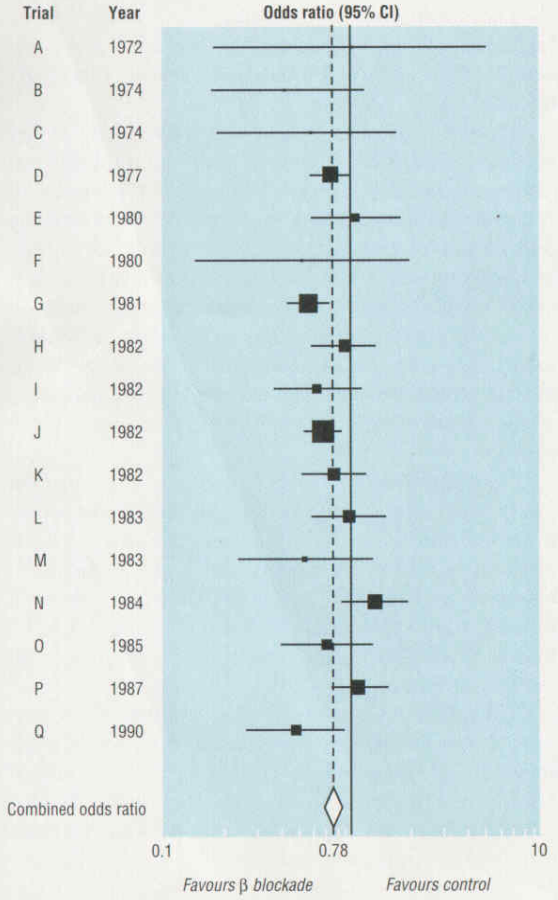


Fig 1 Total mortality from trials of β blockers in secondary prevention after myocardial infarction. The black square and horizontal line correspond to odds ratio and 95% confidence interval for each trial. The size of the black square reflects the weight of each trial. The diamond represents the combined odds ratio and 95% confidence interval, showing 22% a reduction in the odds of death (references are available from the authors)

test for heterogeneity across studies. In such procedures, whether the results of a study reflect a single underlying effect, rather than a distribution of effects, is assessed. If this test shows homogeneous results then the differences between studies are assumed to be a consequence of sampling variation, and a fixed effects model is appropriate. If, however, the test shows that significant heterogeneity exists between study results then a random effects model is advocated. A major limitation with this approach is that the statistical tests lack power—they often fail to reject the null hypothesis of homogeneous results even if substantial differences between studies exist. Although there is no statistical solution to this issue, heterogeneity between study results should not be seen as purely a problem for meta-analysis—it also provides an opportunity for examining why treatment effects differ in different circumstances. Heterogeneity should not simply be ignored after a statistical test is applied; rather, it should be scrutinised, with an attempt to explain it.[24]

Graphic display

Results from each trial are usefully graphically displayed, together with their confidence intervals. Figure 1 represents a meta-analysis of 17 trials of β blockers in secondary prevention after myocardial infarction. Each study is represented by a black square and a horizontal line, which correspond to the point estimate and the 95% confidence intervals of the odds ratio. The 95% confidence intervals would contain the true underlying effect in 95% of the occasions if the study was repeated again and again. The solid vertical line corresponds to no effect of treatment (odds ratio 1.0). If the confidence interval includes 1, then the difference in the effect of experimental and control treatment is not significant at conventional levels ($P > 0.05$). The area of the black squares reflects the weight of the study in the meta-analysis. The confidence interval of all but two studies cross this line, indicating that the effect estimates were non-significant ($P > 0.05$).

The diamond represents the combined odds ratio, calculated using a fixed effects model, with its 95% confidence interval. The combined odds ratio shows that oral β blockade starting a few days to a few weeks after the acute phase reduces subsequent mortality by an estimated 22% (odds ratio 0.78; 95% confidence interval 0.71 to 0.87). A dashed line is plotted vertically through the combined odds ratio. This line crosses the horizontal lines of all individual studies except one (N). This indicates a fairly homogenous set of studies. Indeed, the test for heterogeneity gives a non-significant P value of 0.2.

A logarithmic scale was used for plotting the odds ratios in figure 1. There are several reasons that ratio measures are best plotted on logarithmic scales.[25] Most importantly, the value of an odds ratio and its reciprocal—for example, 0.5 and 2—which represent odds ratios of the same magnitude but opposite directions, will be equidistant from 1.0. Studies with odds ratios below and above 1.0 will take up equal space on the graph and thus look equally important. Also, confidence intervals will be symmetrical around the point estimate.

Relative and absolute measures of effect

Repeating the analysis by using relative risk instead of the odds ratio gives an overall relative risk of 0.80 (95% confidence interval 0.73 to 0.88). The odds ratio is thus close to the relative risk, as expected when the outcome is relatively uncommon (see box). The relative risk reduction, obtained by subtracting the relative risk from 1 and expressing the result as a percentage, is 20% (12% to 27%). The relative measures used in this analysis ignore the absolute underlying risk. The risk of death among patients who have survived the acute phase of myocardial infarction, however, varies widely.[26] For example, among patients with three or more cardiac risk factors the probability of death at two years after discharge ranged from 24% to 60%.[26] Conversely, two year mortality among patients with no risk factors was less than 3%. The absolute risk reduction or risk difference reflects both the underlying risk without treatment and the risk reduction associated with treatment. Taking the reciprocal of the risk difference gives the "number needed to treat" (the number of patients needed to be treated to prevent one event).[17]

For a baseline risk of 1% a year, the absolute risk difference shows that two deaths are prevented per 1000 patients treated (table). This corresponds to 500 patients ($1 \div 0.002$) treated for one year to prevent one

β Blockade in secondary prevention after myocardial infarction—absolute risk reductions and numbers needed to treat for one year to prevent one death for different levels of mortality in control group

One year mortality risk among controls (%)	Absolute risk reduction	No needed to treat
1	0.002	500
3	0.006	167
5	0.01	100
10	0.02	50
20	0.04	25
30	0.06	17
40	0.08	13
50	0.1	10

Calculations assume a constant relative risk reduction of 20%.

death. Conversely, if the risk is above 10%, less than 50 patients have to be treated to prevent one death. Many clinicians would probably decide not to treat patients at very low risk, given the large number of patients that have to be exposed to the adverse effects of β blockade to prevent one death. Appraising the number needed to treat from a patient's estimated risk without treatment and the relative risk reduction with treatment is a helpful aid when making a decision in an individual patient. A nomogram that facilitates calculation of the number needed to treat at the bedside has recently been published.[27]

Meta-analysis using absolute effect measures such as the risk difference may be useful to illustrate the range of absolute effects across studies. The combined risk difference (and the number needed to treat calculated from it) will, however, be essentially determined by the number and size of trials in patients at low, intermediate, or high risk. Combined results will thus be applicable only to patients at levels of risk corresponding to the average risk of the trials included. It is therefore generally more meaningful to use relative effect measures for summarising the evidence and absolute measures for applying it to a concrete clinical or public health situation.

Sensitivity analysis

Opinions will often diverge on the correct method for performing a particular meta-analysis. The robustness of the findings to different assumptions should therefore always be examined in a thorough sensitivity analysis. This is illustrated in figure 2 for the meta-analysis of β blockade after myocardial infarction. Firstly, the overall effect was calculated by different statistical methods, by using both a fixed and a random effects model. The figure shows that the overall estimates are virtually identical and that confidence intervals are only slightly wider with the random effects model. This is explained by the relatively small amount of variation between trials in this meta-analysis.

Secondly, methodological quality was assessed in terms of how patients were allocated to active treatment or control groups, how outcome was assessed, and how the data were analysed.[6] The maximum credit of nine points was given if patient allocation was truly random, if assessment of vital status was independent of treatment group, and if data from all patients initially included were analysed according to the intention to treat principle. Figure 2 shows that the three low quality studies (≤7 points) showed more benefit than the high quality trials. Exclusion of these three studies, however, leaves the overall effect and the confidence intervals practically unchanged.

Thirdly, significant results are more likely to get published than non-significant findings,[28] and this can distort the findings of meta-analyses. The presence of such publication bias can be identified by stratifying the analysis by study size—smaller effects can be significant in larger studies. If publication bias is present, it is expected that, of published studies, the largest ones will report the smallest effects. Figure 2 shows that this is indeed the case, with the smallest trials (50 or fewer deaths) showing the largest effect. However, exclusion of the smallest studies has little effect on the overall estimate.

Finally, two studies (J and N; see fig 1) were stopped earlier than anticipated on the grounds of the results from interim analyses. Estimates of treatment effects from trials that were stopped early are liable to be biased away from the null value. Bias may thus be introduced in a meta-analysis that includes such trials.[29] Exclusion of these trials, however, affects the overall estimate only marginally.

The sensitivity analysis thus shows that the results from this meta-analysis are robust to the choice of the statistical method and to the exclusion of trials of poorer quality or of studies stopped early. It also suggests that publication bias is unlikely to have distorted its findings.

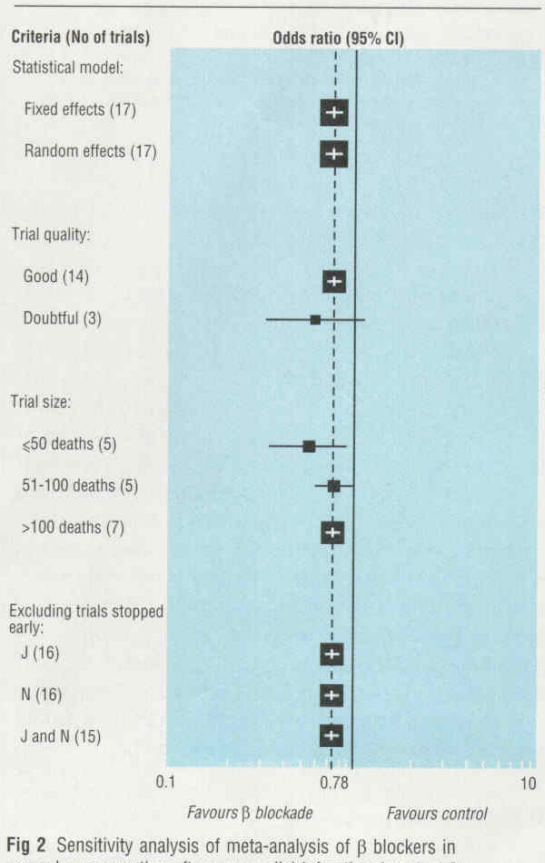


Fig 2 Sensitivity analysis of meta-analysis of β blockers in secondary prevention after myocardial infarction (see text for explanation)

Conclusions

Meta-analysis should be seen as structuring the processes through which a thorough review of previous research is carried out. The issues of completeness and combinability of evidence, which need to be considered in any review,[30] are made explicit. Was it sensible to have combined the individual trials that comprise the meta-analysis? How robust is the result to changes in assumptions? Does the conclusion reached make clinical and pathophysiological sense? Finally, has the analysis contributed to the process of making rational decisions about the management of patients? It is these issues that we explore further in later articles in this series.

The department of social medicine at the University of Bristol and the department of primary care and population sciences at the Royal Free Hospital School of Medicine, London, are part of the Medical Research Council's health services research collaboration.

Funding: ME was supported by the Swiss National Science Foundation.

1 Huque MF. Experiences with meta-analysis in NDA submissions. *Proceedings of the Biopharmaceutical Section of the American Statistical Association* 1988;2:28-33.
2 Egger M, Davey Smith G. Meta-analysis: potentials and promise. *BMJ* 1997;315:1371-4.
3 Egger M, Davey Smith G, Schneider M, Minder CE. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.
4 Sacks H, Chalmers TC, Smith HJ. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233-40.
5 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.
6 Prendiville W, Elbourne D, Chalmers I. The effects of routine oxytocic administration in the management of the third stage of labour: an overview of the evidence from controlled trials. *Br J Obstet Gynaecol* 1988;95:3-16.
7 Begg CB, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:637-9.
8 Greenhalgh T. The Medline database. *BMJ* 1997;315:180-3.
9 Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of Medline searching with a perinatal clinical trial data base. *Controlled Clinical Trials* 1985; 6:306-317.
10 Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994;309:1286-91.
11 Chalmers I, Dickersin K, Chalmers TC. Getting to grips with Archie Cochrane's agenda. *BMJ* 1992;305:786-8.
12 The Cochrane Controlled Trials Register. In: *Cochrane Library*. CD ROM and online. Cochrane Collaboration (issue 1). Oxford: Update Software, 1997.
13 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials* 1995; 16:62-73.
14 Jadad AR, Moore RA, Carrol D, Jenkinson C, Reynolds DJM, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials* 1996; 17:1-12.
15 Chalmers TC. Problems induced by meta-analyses. *Stat Med* 1991;10:971-80.
16 Moher D, Fortin P, Jadad AR, Jüni P, Klassen T, Le Lorier J, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996;347:363-6.
17 Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New Engl J Med* 1988;318:1728-33.
18 Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989;8:141-51.
19 Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985;17:335-71.
20 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986;7:177-88.
21 Carlin JB. Meta-analysis for 2x2 tables: a bayesian approach. *Stat Med* 1992;11:141-58.
22 Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;313:603-7.
23 Eddy DM, Hasselblad V, Shachter R. *Meta-analysis by the confidence profile method. The statistical synthesis of evidence.* Boston: Academic Press, 1992.
24 Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? *Stat Med* 1987;6:351-8.
25 Galbraith R. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889-94.
26 Multicenter Postinfarction Research Group. Risk stratification and survival after myocardial infarction. *New Engl J Med* 1983;309:331-6.
27 Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. *BMJ* 1996;312:426-9.
28 Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867-72.
29 Green S, Fleming TR, Emerson S. Effects on overviews of early stopping rules for clinical trials. *Stat Med* 1987;6:361-7.
30 Oxman AD. Checklists for review articles. *BMJ* 1994;309:648-51.

Words to the wise

Turning and turning in the widening gyre

Cerebral *gyri* and nasal *turbinates* do not, at first, appear to have much in common, but both derive their names from turning words.

The scroll shaped edge of a turbinate bone recalls the spiral structure of a turbinate seashell, which in turn resembles a Roman spinning top, *turbo*. This word was also used for a whirlwind, explaining the connection to *turbine*. The rotary chaos of the whirlwind also explains the Latin *turba*, a disorderly crowd, which gives us *turbulent* and *turbid*, as well as *perturb* and *disturb*.

The cerebral gyri are named for their curved shape: *gyrus* is Latin for a circle or ring, and gives us our word *gyrate*. In 1617 Italian citizens were introduced to the edible root of a recently imported North American sunflower. The taste was a little reminiscent of an artichoke, and so they named the plant the "sunflower artichoke." The Italian word for sunflower is *girasole*, "turn to the sun," and English speakers picked up this word and ran with it, albeit in the wrong direction, so that we now call the same plant a *Jerusalem* artichoke. A couple of hundred years later, in a neighbouring country, Léon Foucault built a large flywheel as a successor to his famous pendulum. When spinning, it maintained its orientation as the earth turned beneath it. So he called it a *gyroscope*, because he could see the earth's rotation if he watched for long enough.

Latin *vertere*, to turn, produces a crop of handy words.Turning the plough at the end of the field gave the Romans *versus*, a furrow, a word they also applied to a line of text. Our own word *verse* comes from this source. *Vertex*, *vortex*, and *vertigo* all derive from the notion of turning around an axis. Watching the sky pass overhead each night inspired the idea of the *universe*, which apparently rotated as a unit. And it is nowadays worth considering that *university* has the same derivation: originally a group of people with one purpose, who behaved as a single entity.

From Latin, too, comes *torquere*, to twist, which gives us *torque* and also *torch*, from the twisted straw that was burnt for illumination. *Torticollis* is, of course, a twisted neck, while *extortion*, *torture*, and *torment* derive their names from the twisting of limbs. *Retort* signifies "twisting back," either of the spoken word or of the neck of a piece of glassware. The legal term *tort* uses twistedness as a metaphor for wrongness; in this case, the wrong caused by a failure of duty. But long before this legalism came into use non-Latin speakers had made precisely the same metaphorical linkage: our English words *wring* and *wrong* can be traced to a common root in the Germanic tongues of ancient northern Europe.

Grant Hutchison, *consultant anaesthetist, Dundee*

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.