The search for evidence of effective health promotion

Viv Speller, Alyson Learmonth, D Harrison

A conceptually sound evidence base for interventions that aim to promote health is urgently required. However, the current search for evidence of effective health promotion is unlikely to succeed and may result in drawing false conclusions about health promotion practice to the long term detriment of public health. The reasons for this are threefold: lack of consensus about the nature of health promotion activity; lack of agreement over what evidence to use to assess effectiveness; and divergent views on appropriate methods for reviewing effectiveness. As a consequence health promotion may be designated "not effective" because it is being assessed with inappropriate tools.

What are we looking for?

Health promotion is a multifactorial process operating on individuals and communities, through education, prevention, and protection measures.[1] The statement of principles known as the Ottawa charter for health promotion, developed by the World Health Organisation, is internationally accepted as the guiding framework for health promotion activity. This describes five approaches; building healthy public policy, creating supportive environments, strengthening community action, developing personal skills, and reorienting health services.[2] Health promotion methods may include activities as diverse as awareness raising campaigns, provision of health information and advice, influencing social policy, lobbying for change, professional training, and community development—often in combination in complex interventions. However, health promotion is rarely judged on its effectiveness in all these areas.

Not just about individual behaviour

In Britain the Health of the Nation strategy's concentration on reducing disease and behavioural risk factors[3] has overemphasised the role of health promotion in developing personal knowledge and skills and focused attention on assessing individual health outcomes. In Europe, however, health promotion emphasises the development of health promoting settings, such as schools, workplaces and hospitals, which aim to enable and support healthy behaviour. Practice therefore includes management and organisational development approaches.[4] In the Pacific countries the emphasis is on sustaining healthy environments, and legislative approaches are aimed at populations, rather than individuals.[5]

Summary points

Health promotion in the UK is at risk from the application of inappropriate methods of assessing evidence, overemphasis on outcomes of individual behaviour change, and pressure on resources.

The effect of health promotion should be assessed across the breadth of its activities and settings, not just by changes in individual health behaviour

Systematic review methods need to be revised to include a broader range of studies and research methods, including qualitative research

Review criteria should consider the quality of the health promotion intervention as well as the research design

Seeking evidence of change in individual behaviour or health outcome is not appropriate if interventions are aiming to achieve other types of change. It may be difficult to generalise from studies conducted on different populations because the interventions and research methods may have been chosen to answer different questions about different types of health promotion practice. The logic of looking for immediate changes in health or behaviour as a result of one shot interventions, such as advice from health professionals or advertising, is questionable when we know that attitudes and behaviours are shaped by a panoply of socioeconomic and cultural influences. Attribution of any changes detected to a single intervention is also dubious. Despite calls to refocus population health "upstream," many research studies are still seeking such simple results with little consideration of the effect of other influences, or use costly designs to attempt to control them.[6]

How should we look for evidence of effectiveness?

Britain's national research and development programme is not supporting health promotion research because it is dominated by clinically oriented criteria and methods. Evidence based medicine is defined as

Wessex Institute for Health Research and Development, Winchester SO22 5DH
Viv Speller, *senior lecturer*

North Durham Community Health Care Trust, Health Centre, Chester-le-Street, Durham DH3 3UR
Alyson Learmonth, *director of health promotion*

North West Lancashire Health Promotion Unit, Sharoe Green Hospital, Preston PR2 8DU
D Harrison, *health promotion general manager*

Correspondence to: Dr Speller.

BMJ 1997;315:361–3

CHRISTOPHE BLUNTZER/IMPACT

There is more to health promotion than reducing behavioural risk factors

"the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."[7] The systematic review process aims to ensure reliable and rigorous evaluation of evidence. Criteria for the inclusion of studies usually place randomised controlled trials as the gold standard for judging whether a treatment is effective. This approach has been directly transferred to health promotion in the series of reviews by the NHS Centre for Reviews and Dissemination to determine what is known about the effectiveness of health promotion interventions. This evidence is now becoming available through the *Effective Health Care Bulletin* series[8 9] and the Health Education Authority series of health promotion effectiveness reviews.[10 11]

The selection of studies for inclusion is done on the basis of the quality of the research only, not on the quality of the health promotion intervention. This can produce some anomalous results. An earlier bulletin on brief interventions and alcohol use[12] suggested that brief interventions are as effective as more expensive specialist treatment. This sparked a debate about the meaning of the term "brief interventions" which highlighted the risks inherent in glossing over the detail.[13] The intervention techniques had been pooled because they were all said to be of similar short duration and had common characteristics. However, these ranged from five minutes of simple advice to regular sessions over six months of structured interventions by a general practitioner. It is clearly difficult from this for the commissioner or practitioner to determine which intervention technique is most successful.

Another systematic review looked at the effectiveness of sexual health education interventions for young people.[14] Of 270 papers reporting sexual health interventions only 12 met the inclusion criteria. Criteria such as the appropriateness of the interventions studied and outcome measures used were not considered essential "because of the large element of subjectivity in assessing whether they had been met." This can lead to spurious generalised conclusions, such as that drawn from a US study of an abstinence education programme for 13 year old boys from a low income minority group, that chastity education is harmful. The study was included despite high attrition rates and dependence on self reported outcomes. After six lessons aimed at reducing premarital sex, more of the intervention group claimed to have initiated sexual intercourse. An experienced health promoter would immediately know that this is unlikely to have been an appropriate intervention method for this target group.

Process evaluation may be useful

As randomised trials are expensive, it is important that they are applied only to study high quality interventions that have been developed appropriately and based on knowledge of best practice. Another approach to assessing effectiveness which pays more attention to the quality of the intervention has been attempted by the International Union for Health Promotion and Education in a series of 16 effectiveness reviews.[15] Criteria for selecting studies included information about the strategies used in the intervention. Evaluation criteria included not only the use of controls and measurements before and after the intervention but also formative or process evaluation. This is the study of the processes of implementing the intervention, to answer such questions as: was it applied in the manner intended, did other factors come into play that might have affected the result, what did the participants think about the process? Process evaluation often uses qualitative research methods and complements outcome evaluation. In health promotion research the technique of "triangulation"—that is, drawing conclusions from a number of different sources of data—is also used.

There are several important differences between the approach to review taken by the International Union for Health Promotion and Education and the systematic review approach of the Centre for Reviews and Dissemination. While the latter pays inadequate attention to the process of the intervention, the former is insufficiently critical of the soundness of quantitative research methods. It would be helpful to attempt to combine the best elements of both for future effectiveness reviews.

Another fundamental problem with using randomised controlled trials in health promotion research is that where interventions aim to influence systems or populations it may be difficult to randomly allocate units such as schools or communities to intervention or control groups, so quasiexperimental control designs are used. Well known quasiexperimental studies include the Stanford heart disease prevention program[16] and the Minnesota heart health program,[17] where multiple interventions were applied to communities and their risk factor profiles were subsequently compared with matched comparison communities. One of the major problems with studies employing this design is the "contamination" of the control group. This poses a serious dilemma in that the practice of health promotion relies strongly on the diffusion of the effects of the intervention through the target community. The difficulty of controlling for spillover to the comparison community reduces the effect attributable to the intervention. Whether health gain in the control group is influenced by diffusion of the intervention, or the intervention group effects are produced by secular trends, cannot be determined by looking at outcome measures alone.

As has been recognised for healthcare evaluation, a wide range of research methods is warranted.[18] Qualitative research can contribute to assessing the effectiveness of interventions by illuminating processes, exploring diversity, and developing new theories. It includes a broad range of methods such as case study, ethnography, action research, participant observation, conversation analysis, and grounded theory.[19] For example, a textual analysis of the way health visitors provide information to first time mothers identified that they frequently failed to reinforce the mother as a skilful and knowledgeable person, thereby affecting the reception of advice.[20] Ethnographic studies provide a way of assessing outcomes of health promotion interventions on activities such as injecting drug use and cannabis dealing which may not be amenable to other research methods.[21 22]

Will we recognise effective health promotion when we find it?

An inquiry by Britain's parliamentary public accounts committee into the cost effectiveness of the Health of the Nation strategy showed that spending on health promotion in 1996 was £3m on the strategy, £45m on health education via the Health Education Authority, £73m on paying general practitioners for health promotion work, and £90m on NHS health promotion units.[23] This represents less than 1% of the NHS's annual budget and less than the expenditure on staff cars and travelling and subsistence in 1994-5.[24] Yet anecdotal evidence suggests that cost pressures, coupled with the inability to present conclusive evidence of effectiveness, are conspiring to make health promotion contracts a soft option for budget cuts in next year's contracts.

To get out of this downward spiral health promotion workers must demonstrate evidence of its effectiveness by establishing a robust evidence base. Existing reviews should be critically reanalysed using appropriate inclusion criteria which consider the quality of the health promotion intervention as well as that of the research. Criteria for rigorous evaluation of qualitative research methods need to be derived. The search should be broadened to include studies that measure the impact of interventions on systems and organisational development as well as change in individual behaviour. Further work needs to be done using the existing health topic based reviews to analyse cross cutting themes to examine the effectiveness of different methods of health promotion.

ZEFA

But will this make her healthier?

Research should be commissioned to fill the gaps identified by the reviews, ensuring that an appropriate range of methods is used. In doing this practitioners need to be involved in designing and implementing viable interventions including collecting data for process evaluation to ensure that programmes function optimally. It is essential for health promotion research to bring together experts from a range of disciplines to design and conduct studies that pay adequate attention to both the quality of the intervention and the methods of evaluation.

Finally, the evidence base must be accessible to and used by practitioners. While practitioners need to be more critical and to substantiate their decisions with evidence, key messages must also be disseminated clearly and unequivocally to influence practice.

If the commitment enshrined in the mission for the NHS—to promote health and prevent disease as well as providing treatment—is to be taken seriously, then the national research and development agenda needs to consider ways of tackling these issues. Action needs to be taken urgently to redress the negative consequences on health promotion of a misdirected search which is veering off course.

1 Tannahill A. What is health promotion? *Health Education Journal* 1985;44:167-8.

2 World Health Organisation. *The Ottawa charter: principles for health promotion.* Copenhagen: WHO Regional Office for Europe, 1986.

3 Department of Health. *Health of the nation.* London: HMSO, 1992.

4 Grossman R, Scala K. *Health promotion and organisational development: developing settings for health.* Vienna: WHO, 1993.

5 Central Sydney Area Health Service and NSW Health. *Program management guidelines for health promotion.* Sydney: Better Health Centre, 1994.

6 Population health looking upstream. *Lancet* 1994;343:429-30.

7 Sackett DL, Rosenberg WC, Muir Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312:71-2.

8 NHS Centre for Reviews and Dissemination. Preventing falls and unintentional injuries in older people. *Effective Health Care* 1996;2:4.

9 NHS Centre for Reviews and Dissemination. Unintentional injuries in young people. *Effective Health Care* 1996;2:5.

10 Ebrahim S, Davey-Smith G. Health promotion in older people for the prevention of coronary heart disease and stroke. London: Health Education Authority, 1996.

11 Towner E, Dowswell T, Simpson G, Jarvis S. Health promotion in childhood and young adolescence for the prevention of unintentional injuries. London: Health Education Authority, 1996.

12 NHS Centre for Reviews and Dissemination. Brief interventions and alcohol use. *Effective Health Care* 1993;7.

13 Heather N. Interpreting the evidence on brief interventions for excessive drinkers: the need for caution. *Alcohol and Alcoholism* 1995;30:287-96.

14 Oakley A, Fullerton D, Holland J, Arnold S, France-Dawson M, Kelley P, et al. Sexual health interventions for young people: a methodological review. *BMJ* 1995;310,158-62.

15 Veen CA, Vereijken I, van Driel WG, Belien M A. *An instrument for analysing effectiveness studies on health promotion and health education.* Utrecht: Dutch Centre for Health Promotion and Health Education and IUHPE/EURO 1994.

16 Farquhar J, Fortmann S, Flora J, Taylor CB, Haskell WL, Williams PT, et al. Effects of community-wide education on cardiovascular disease risk factors. The Stanford five city project. *JAMA* 1990;264:359-65.

17 Luepker R, Murray D, Jacobs D, Mittelmark MB, Bracht N, Carlaw R, et al. Community education for cardiovascular disease prevention: risk factor changes in the Minnesota heart health program. *Am J Pub Health* 1994;84:1383-93.

18 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.

19 Denzin NK, Lincoln YS. *Handbook of qualitative research.* London: Sage, 1994.

20 Heritage J, Sefi S. Dilemmas of advice giving: aspects of the delivery and reception of advice in interactions between health visitors and first time mothers. In: Drew P, Heritage J, eds. *Talk at work.* Cambridge: Cambridge University Press, 1992:59-417.

21 Taylor A. *Women drug users: an ethnography of a female injecting community.* Oxford: Clarendon Press, 1993.

22 Fountain J. Dealing with data. In: Hobbs D, May T, eds. *Interpreting the field.* Beverley Hills, CA: Sage, 1993:145-73.

23 Limb L. Health of the Nation under scrutiny. *Health Services Journal* 1996;14 Nov:8.

24 Monitor. *Health Services Journal* 1996;14 Nov:25.

(Accepted 3 February 1997)

How to read a paper

Statistics for the non-statistician. I: Different types of data need different statistical tests

This is the fourth in a series of 10 articles introducing non-experts to finding medical articles and assessing their value

Unit for Evidence-Based Practice and Policy, Department of Primary Care and Population Sciences, University College London Medical School/ Royal Free Hospital School of Medicine, Whittington Hospital, London N19 5NF

Trisha Greenhalgh, *senior lecturer*

p.greenhalgh@ucl.ac.uk

BMJ 1997;315:364–6

As medicine leans increasingly on mathematics no clinician can afford to leave the statistical aspects of a paper to the "experts." If you are numerate, try the "Basic Statistics for Clinicians" series in the *Canadian Medical Association Journal*,[1-4] or a more mainstream statistical textbook.[5] If, on the other hand, you find statistics impossibly difficult, this article and the next in this series give a checklist of preliminary questions to help you appraise the statistical validity of a paper.

Have the authors set the scene correctly?

Have they determined whether their groups are comparable, and, if necessary, adjusted for baseline differences?

Most comparative clinical trials include either a table or a paragraph in the text showing the baseline characteristics of the groups being studied. Such a table should show that the intervention and control groups are similar in terms of age and sex distribution and key prognostic variables (such as the average size of a cancerous lump). Important differences in these characteristics, even if due to chance, can pose a challenge to your interpretation of results. In this situation, adjustments can be made to allow for these differences and hence strengthen the argument.[6]

What sort of data have they got, and have they used appropriate statistical tests?

Numbers are often used to label the properties of things. We can assign a number to represent our height, weight, and so on. For properties like these, the measurements can be treated as actual numbers. We can, for example, calculate the average weight and height of a group of people by averaging the measurements. But consider an example in which we use numbers to label the property "city of origin," where 1 = London, 2 = Manchester, 3 = Birmingham, and so on. We could still calculate the average of these numbers for a particular sample of cases, but we would be completely unable to interpret the result. The same would apply if we labelled the property "liking for *x*" with 1 = not at all, 2 = a bit, and 3 = a lot. Again, we could calculate the "average liking," but the numerical result would be uninterpretable unless we knew that the difference between "not at all" and "a bit" was exactly the same as the difference between "a bit" and "a lot."

PETER BROWN

Summary points

In assessing the choice of statistical tests in a paper, first consider whether groups were analysed for their comparability at baseline

Does the test chosen reflect the type of data analysed (parametric or non-parametric, paired or unpaired)?

Has a two tailed test been performed whenever the effect of an intervention could conceivably be a negative one?

Have the data been analysed according to the original study protocol?

If obscure tests have been used, do the authors justify their choice and provide a reference?

All statistical tests are either parametric (that is, they assume that the data were sampled from a particular form of distribution, such as a normal distribution) or non-parametric (they make no such assumption). In general, parametric tests are more powerful than non-parametric ones and so should be used if possible.

Non-parametric tests look at the rank order of the values (which one is the smallest, which one comes next, and so on) and ignore the absolute differences between them. As you might imagine, statistical significance is more difficult to show with non-parametric tests, and this tempts researchers to use statistics such as the *r* value inappropriately. Not only is the *r* value (parametric) easier to calculate than its non-parametric equivalent but it is also much more likely to give (apparently) significant results. Unfortunately, it will give a spurious estimate of the significance of the result, unless the data are appropriate to the test being used. More examples of parametric tests and their non-parametric equivalents are given in table 1.

Another consideration is the shape of the distribution from which the data were sampled. When I was at school, my class plotted the amount of pocket money received against the number of children receiving that amount. The results formed a histogram the same shape as figure 1—a "normal" distribution. (The term "normal" refers to the shape of the graph and is used

Table 1 Some commonly used statistical tests

Parametric test	Example of equivalent non-parametric test	Purpose of test	Example
Two sample (unpaired) *t* test	Mann-Whitney U test	Compares two independent samples drawn from the same population	To compare girls' heights with boys' heights
One sample (paired) *t* test	Wilcoxon matched pairs test	Compares two sets of observations on a single sample	To compare weight of infants before and after a feed
One way analysis of variance (*F* test) using total sum of squares	Kruskall-Wallis analysis of variance by ranks	Effectively, a generalisation of the paired *t* or Wilcoxon matched pairs test where three or more sets of observations are made on a single sample	To determine whether plasma glucose level is higher one hour, two hours, or three hours after a meal
Two way analysis of variance	Two way analysis of variance by ranks	As above, but tests the influence (and interaction) of two different covariates	In the above example, to determine if the results differ in male and female subjects
χ^2 test	Fisher's exact test	Tests the null hypothesis that the distribution of a discontinuous variable is the same in two (or more) independent samples	To assess whether acceptance into medical school is more likely if the applicant was born in Britain
Product moment correlation coefficient (Pearson's *r*)	Spearman's rank correlation coefficient (r_σ)	Assesses the strength of the straight line association between two continuous variables.	To assess whether and to what extent plasma HbA_1 concentration is related to plasma triglyceride concentration in diabetic patients
Regression by least squares method	Non-parametric regression (various tests)	Describes the numerical relation between two quantitative variables, allowing one value to be predicted from the other	To see how peak expiratory flow rate varies with height
Multiple regression by least squares method	Non-parametric regression (various tests)	Describes the numerical relation between a dependent variable and several predictor variables (covariates)	To determine whether and to what extent a person's age, body fat, and sodium intake determine their blood pressure

because many biological phenomena show this pattern of distribution). Some biological variables such as body weight show "skew normal" distribution, as shown in figure 2. (Figure 2 shows a negative skew, whereas body weight would be positively skewed. The average adult male body weight is 70 kg, and people exist who weigh 140 kg, but nobody weighs less than nothing, so the graph cannot possibly be symmetrical.)

Non-normal (skewed) data can sometimes be transformed to give a graph of normal shape by performing some mathematical transformation (such as using the variable's logarithm, square root, or reciprocal). Some data, however, cannot be transformed into a smooth pattern. For a very readable discussion of the normal distribution see chapter 7 of Martin Bland's *Introduction to Medical Statistics.*[5]

Deciding whether data are normally distributed is not an academic exercise, since it will determine what type of statistical tests to use. For example, linear regression will give misleading results unless the points on the scatter graph form a particular distribution about the regression line—that is, the residuals (the perpendicular distance from each point to the line) should themselves be normally distributed. Transforming data to achieve a normal distribution (if this is indeed achievable) is not cheating: it simply ensures that data values are given appropriate emphasis in assessing the overall effect. Using tests based on the normal distribution to analyse non-normally distributed data, however, is definitely cheating.

If the authors have used obscure statistical tests, why have they done so and have they referenced them?
The number of possible statistical tests sometimes seems infinite. In fact, most statisticians could survive with a formulary of about a dozen. The rest should generally be reserved for special indications. If the paper you are reading seems to describe a standard set of data which have been collected in a standard way, but the test used has an unpronounceable name and is not listed in a basic statistics textbook, you should smell a rat. The authors should, in such circumstances, state why they have used this test, and give a reference (with page numbers) for a definitive description of it.

Are the data analysed according to the original protocol?
If you play coin toss with someone, no matter how far you fall behind, there will come a time when you are one ahead. Most people would agree that to stop the game then would not be a fair way to play. So it is with research. If you make it inevitable that you will (eventually) get an apparently positive result you will also make it inevitable that you will be misleading yourself about the justice of your case.[7] (Terminating an intervention trial prematurely for ethical reasons when subjects in one arm are faring particularly badly is a different matter and is discussed elsewhere.[7])

Raking over your data for "interesting results" (retrospective subgroup analysis) can lead to false conclu-

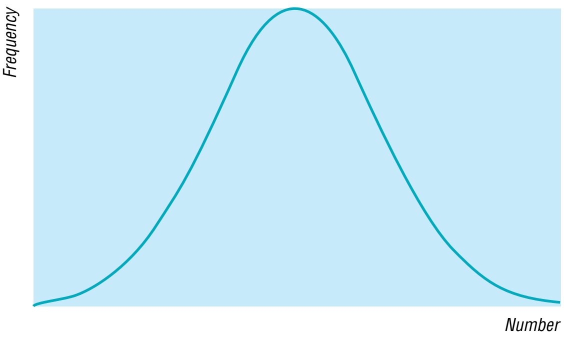

Fig 1 Normal curve

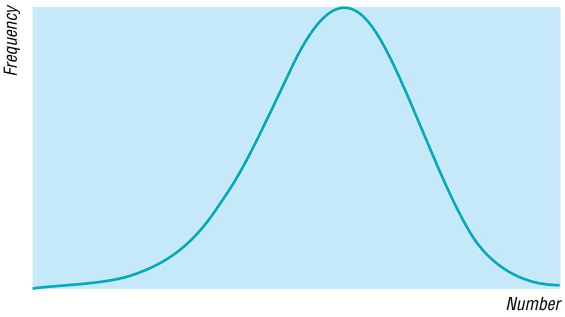

Fig 2 Skewed curve

sions.[8] In an early study on the use of aspirin in preventing stroke, the results showed a significant effect in both sexes combined, and a retrospective subgroup analysis seemed to show that the effect was confined to men.[9] This conclusion led to aspirin being withheld from women for many years, until the results of other studies[10] showed that this subgroup effect was spurious.

This and other examples are included in Oxman and Guyatt's, "A consumer's guide to subgroup analysis," which reproduces a useful checklist for deciding whether apparent subgroup differences are real.[11]

Paired data, tails, and outliers

Were paired tests performed on paired data?

Students often find it difficult to decide whether to use a paired or unpaired statistical test to analyse their data. There is no great mystery about this. If you measure something twice on each subject—for example, blood pressure measured when the subject is lying and when standing—you will probably be interested not just in the average difference of lying versus standing blood pressure in the entire sample, but in how much each individual's blood pressure changes with position. In this situation, you have what is called "paired" data, because each measurement beforehand is paired with a measurement afterwards.

In this example, it is using the same person on both occasions which makes the pairings, but there are other possibilities (for example, any two measurements of bed occupancy made of the same hospital ward). In these situations, it is likely that the two sets of values will be significantly correlated (for example, my blood pressure next week is likely to be closer to my own blood pressure last week than to the blood pressure of a randomly selected adult last week). In other words, we would expect two randomly selected paired values to be closer to each other than two randomly selected unpaired values. Unless we allow for this, by carrying out the appropriate paired sample tests, we can end up with a biased estimate of the significance of our results.

Was a two tailed test performed whenever the effect of an intervention could conceivably be a negative one?

The term "tail" refers to the extremes of the distribution—the areas at the outer edges of the bell in figure 1. Let's say that the graph represents the diastolic blood pressures of a group of people of which a random sample are about to be put on a low sodium diet. If a low sodium diet has a significant lowering effect on blood pressure, subsequent blood pressure measurements on these subjects would be more likely to lie within the left tail of the graph. Hence we would analyse the data with statistical tests designed to show whether unusually low readings in this patient sample were likely to have arisen by chance.

But on what grounds may we assume that a low sodium diet could only conceivably put blood pressure down, but could never do the reverse, put it up? Even if there are valid physiological reasons in this particular example, it is certainly not good science always to assume that you know the direction of the effect which your intervention will have. A new drug intended to relieve nausea might actually exacerbate it, or an educational leaflet intended to reduce anxiety might increase it. Hence, your statistical analysis should, in general, test the hypothesis that either high or low values in your dataset have arisen by chance. In the language of the statisticians, this means you need a two tailed test, unless you have very convincing evidence that the difference can only be in one direction.

Were "outliers" analysed with both common sense and appropriate statistical adjustments?

Unexpected results may reflect idiosyncrasies in the subject (for example, unusual metabolism), errors in measurement (faulty equipment), errors in interpretation (misreading a meter reading), or errors in calculation (misplaced decimal points). Only the first of these is a "real" result which deserves to be included in the analysis. A result which is many orders of magnitude away from the others is less likely to be genuine, but it may be so. A few years ago, while doing a research project, I measured several different hormones in about 30 subjects. One subject's growth hormone levels came back about 100 times higher than everyone else's. I assumed this was a transcription error, so I moved the decimal point two places to the left. Some weeks later, I met the technician who had analysed the specimens and he asked, "Whatever happened to that chap with acromegaly?"

Statistically correcting for outliers (for example, to modify their effect on the overall result) requires sophisticated analysis and is covered elsewhere.[6]

I am grateful to Mr John Dobby for educating me on statistics and for repeatedly checking and amending this article. Responsibility for any errors is mine alone.

1 Guyatt G, Jaenschke R, Heddle, N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. 1. Hypothesis testing. *Can Med Assoc J* 1995;152:27-32.

2 Guyatt G, Jaenschke R, Heddle, N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. 2. Interpreting study results: confidence intervals. *Can Med Assoc J* 1995;152:169-73.

3 Jaenschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle, N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *Can Med Assoc J* 1995;152:351-7.

4 Guyatt G, Walter S, Shannon H, Cook D, Jaenschke R, Heddle, N. Basic statistics for clinicians. 4. Correlation and regression. *Can Med Assoc J* 1995;152:497-504.

5 Bland M. *An introduction to medical statistics.* Oxford: Oxford University Press, 1987.

6 Altman D. *Practical statistics for medical research.* London: Chapman and Hall, 1995.

7 Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* 1987;7:1231-42.

8 Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. *Int J Health Technology Assessment* 1996;12:264-75.

9 Canadian Cooperative Stroke Group. A randomised trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53-9.

10 Antiplatelet Trialists Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ* 1988;296:320-1.

11 Oxman, AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med* 1992;116:79-84.