A user's guide to evidence-based oncology

Søren M. Bentzen

Gray Cancer Institute, PO Box 100, Mount Vernon Hospital, Northwood, Middlesex HA6 2JR, UK

Introduction

Evidence-based medicine (EBM) has been defined as ". . . the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" [1]. Clearly, this is a compelling concept — just think about the alternative — and this new paradigm for clinical excellence has had an increasing number of followers. At the heart of EBM is the accumulation and critical synthesis of outcome data from clinical studies, in particular from randomised controlled trials. With an increasing push towards medical practice being evidence-based, the quality, or indeed the validity, of this "evidence" obviously becomes crucial.

While the consumers of health care are the patients, the users of medical evidence are primarily the treating physicians. Admittedly, medical evidence is being used, and misused, also by politicians, health care administrators and even the patients themselves. Yet, the ultimate responsibility for the prescription and administration of therapy lies with the physician. Clinical scientists are involved in the generation of "evidence" and this is a challenging task in itself. There are numerous text books and a number of teaching courses targeting the potential clinical trialist. The *users* of this "evidence" are facing an equally challenging task: they must be skilled in appraising the evidence, often based on incomplete or even potentially misleading accounts of the studies that have produced the data in question. In this review, we take the user's perspective and although other study designs may provide useful insights as well, we will focus the discussion on publications of the results of randomised controlled trials.

There are a huge number of national or professional guidelines and recommendations concerning standard care for various patients. While these are most often very helpful, the critical user will have to go back to the original sources of medical evidence and appraise the strength of evidence using his or her personal judgement.

Means of publication of medical research

Medical research findings are disseminated through many different channels.

Oral presentations, posters, published papers

The three main types of primary publication are oral presentations, posters and published papers and although the peer-reviewed published paper is the most definitive publication of a study, oral and poster presentations are also useful sources of information. The strength — and weakness — of the oral presentation is that it is a more personal and less formal mode of communication than a written paper. Most often an oral presentation is a compact account of a piece of work, delivered under severe time constraints, say, in a 10-minute presentation, and therefore with a strict prioritisation of the information that is included. Most researchers have a clear idea of the limitations to their own study, and they will often be more open about these in an oral rather than a written communication. Oral presentations provide a unique opportunity to ask questions to the investigator or challenge various aspects of the study. Unfortunately, the time allotted for discussion at scientific meetings is often not sufficient to allow a detailed discussion; but then of course there are the coffee breaks and virtually all presenters, no matter how distinguished they are, will be more than happy to discuss further aspects of their study at length!

The informality of oral communications is also their weakness. It may be difficult precisely to recall, not to say prove, what was said in a presentation or a discussion and any specific statement may potentially be misleading if cited out of its context. Most of us will occasionally say things in the heat of a discussion that we would not necessarily want to be cited for in print.

Posters are in many ways in between oral presentations and written papers. The context is more structured but there are still severe constraints on the amount of material that can be presented. Remember also that posters are often the first presentation of data from a study and this means that the study or the data analysis may still be in progress at the time of producing the poster. The interpretation of the study findings may be refined when a more definitive analysis is published in written form. Poster presentation sessions where poster authors attend their poster and are available for one-on-one discussion can be particularly helpful for discussing details of a study.

Peer-review: what it is and what it doesn't

Peer-review is an almost sacred principle in the scientific literature and although the system has many flaws it is difficult to see a good alternative. The idea is that when a manuscript is submitted to a journal, the editorial office sends the manuscript to a number of external referees, typically two or three, and asks them for an evaluation of the suitability and the priority to publish this in the journal. These referees are scientists or clinicians who are experts in the relevant field of research. Perhaps it is worth mentioning that the referees are unpaid and that it is quite demanding and time-consuming to write a good review of, in particular a poor, paper. Many journals use a semi-structured review form, asking specifically for a number of issues relating to the quality of the presentation and the originality, validity and importance of the study findings. In addition to this, the referees are asked to provide advice to the authors on possible improvements of the manuscript and to motivate their recommendation as to the acceptability of publishing the manuscript. Clearly, this system involves an element of subjective judgment but in practice the decision to reject or accept a paper is often more obvious than one should think. It does happen that two referees have strongly opposing views on a manuscript and in this case it is the editor, if necessary after asking for a third expert opinion, who will decide the fate of the manuscript.

From the reader's perspective, peer-review is definitely a quality stamp; non-peer-reviewed material should be treated with sound scepticism or even suspicion. However, it is also important to realise the limitations to the peer-review system. Peer-review IS NOT a warranty against the publication of flawed research or poor reporting of studies. Godlee et al. from the *British Medical Journal* [2] introduced 8 deliberate errors or weaknesses into a paper about to be published in the journal and sent it to 420 potential reviewers. Among the 221 respondents (53%), the median number of weaknesses commented on by the reviewers was two. Nobody commented on more than five, and one in six reviewers did not comment on a single one of the 8 errors.

Scientific correspondence

Most scientific journals publish letters to the editor, often commenting on a recently published paper in the journal or perhaps reporting a scientific observation that it is felt does not require the length of a full research paper. The typical time from submission to printing of a letter is shorter than for full papers, which improves the flow of the debate. Depending on journal policy, letters may or may not be peer reviewed. However, it is standard practice that the authors of a paper being commented upon are given the opportunity to write a rebuttal and these exchanges of views may be both entertaining and informative. Literature retrieval systems, such as MEDLINE, contain direct references from a paper to any related published correspondence or editorial comments. It is highly recommended to study these when forming your opinion of a paper.

Internet websites

This is a growing information channel and it is currently somewhat of a minefield: the best web sites are really good and the worst are really bad; but unfortunately it is not always easy to tell the difference. A recent study found that the rank of web sites on "breast cancer" using the Google search engine was related to the type of content of the site [3]. However, the rank was not correlated with measures of quality such as display of authorship, attribution or references, currency of information, and disclosure. Jadad and Gagliardi [4] reviewed 47 proposed scales for rating the quality of medical information on websites and concluded that none of them had been validated and that it was unclear if they were useful at all.

Critical reading

There are two principal rules in critical reading. Rule no. 1: Read the paper! Rule no. 2: Look at the data!

Regarding the first of these rules, it is amazing how often people support a statement or a point of view by referring to a paper *that they have never actually read*! This happens especially in this age of electronic literature retrieval systems like the US National Library of Medicine's MEDLINE database. These systems are extremely powerful: MEDLINE places bibliographical details of more than 12 million papers published in 4,600 journals since the mid-1960s literally at your fingertips. MEDLINE also contains the abstracts of most of these papers and you can read these for free. These databases are key tools for most researchers and they are of tremendous value for the scientific community. Obviously, most of the journals indexed in, say, MEDLINE are not in the user's local library. Some journals offer online access to their contents but most often this is only free to journal subscribers while other potential readers will have to pay-per-view. All of this, together with the sheer amount of work involved in actually reading through the often dozens of potentially interesting papers identified in your literature search, makes it tempting to quote a paper based on reading the abstract only. This is a very dangerous practice since abstracts are often not precise accounts of the real contents of a paper.

The other situation where Rule no. 1 is violated is when primary references are quoted based on a secondary reference such as another research paper or a review. Again, this is a dangerous practice as documented by Evans et al. [5] in a review of the accuracy of quoting 137 randomly chosen references from *The American Journal of Surgery*; *Surgery, Gynecology and Obstetrics*; and *Surgery*. A major error of quotation was assigned if the referenced article failed to substantiate, was unrelated to, or even contradicted the assertion made by the quoting authors. A depressing 27% of the quotations (37 of 137) were classified as major errors. It is probably unrealistic to expect even the keenest reader to check all primary references of a paper. However, in case of key findings quoted in a paper, a closer look at the primary reference is a must for the critical reader.

Rule no. 2 is equally important: you should never rely on the authors' conclusion but try to reach your own based on the data presented in a paper. In the standard lay-out of a scientific paper, the Introduction, Discussion and Conclusion sections are mainly window dressing. Not to say that they cannot be informative if they are well written, but the substance of a paper is in the Methods and Materials and in the Results sections.

A rough guide to what to look for in a report of a randomised controlled trial (RCT) is found in Table 1. This is not meant as a checklist that can be ticked while reading the paper as there is no simple yes/no answer to most of these. It is more like a shopping list of ingredients that the critical reader will need in order to assess the strength of evidence presented in a paper. Some of the items in this table will be discussed later.

Grading the strength of evidence

Clearly, it would be extremely helpful to have a validated instrument for assessing the strength of evidence produced by a given study and this has been the subject of considerable research efforts. A recent systematic review prepared by West and colleagues for the U.S. Department of Health and Human Services found a total of 1602 publications relating to assessment of study quality or strength of evidence [6]. A total of 121 systems were reviewed and these included scales, checklists or other instruments or guidelines for assessment of quality or evidence strength. Twenty of these systems were developed for evaluation of systematic reviews and 49 systems for RCTs. In addition, the report reviewed 40 systems for grading the strength of a body of evidence.

Unfortunately, there is no consensus as to the most adequate quality assessment instrument and although there is a considerable overlap between items included in the various scales or checklists, there are also important differences. This was convincingly shown by Juni et al. [7] who applied 25 different scales to assess the quality of 17 trials comparing low molecular weight heparin (LMWH) with standard heparin in the prevention of post-operative thrombosis. For each scale, the possible relative advantage of LMWH was estimated separately in two strata: high-quality and low-quality trials. For six of the scales, high-quality studies showed no significant difference between the two types of heparin whereas low-quality studies showed a significant benefit from LMWH. For seven other scales, the result was the opposite: high-quality studies showed a benefit from LMWH whereas low-quality studies did not. For the remaining 12 scales, the effect estimates were similar in the two quality strata.

One of the most frequently cited and used of these quality assessment scales is the simple scale proposed by Jadad et al. [8] which assesses three elements of study design: whether the study is appropriately randomised, whether it is appropriately blinded and whether withdrawals and dropouts are described. The resulting score ranges from 0 to 5 (see Table 2). The scale was developed and applied in the evaluation of the quality of 36 reports on RCTs in pain research. The authors used a panel of assessors and showed that this scale produced consistent results and that the effect estimate from double-blind trials was significantly lower that the estimate from non-blinded trials. Subsequent authors have applied the Jadad scale in diverse fields of medicine. There are, however, two obvious but important caveats here. First

Table 1

Key elements in judging the strength of evidence and relevance of a specific randomised controlled trial[a]

Domain	Elements
Study question	Clearly focused and appropriate question
Study design	Description of study population, specific inclusion and exclusion criteria Planned treatments and their timing Target sample size with justification, minimum clinically relevant difference
Randomisation	Method used for randomisation Method of concealment, separation of generator of random treatment assignment from treating physician Similarity of groups at baseline with respect to important demographic or prognostic factors
Blinding	Procedures for blinding (if relevant) of investigators, subjects, assessors, etc.
Interventions	Intervention(s) clearly detailed for all study groups (e.g., dose, route, timing for drugs, details of radiotherapy dose-fractionation, target volume, dose distribution, supportive care) Compliance with intervention Equal care of groups except for intervention
Endpoints and follow-up	Primary and secondary endpoints: Are they relevant? Are any important endpoints missing? Frequency and intensity (diagnostic procedures etc.) of follow-up, number of patients at risk as a function of time Incidence and grade of early and late treatment-related toxicity Trial monitoring, early stopping rules, comparison of actual sample size with target sample size, reasons for deviations, statistical adjustments performed
Statistical analysis	Appropriate statistical analyses, treatment outcome analysed according to intention-to-treat Statistical power, confidence intervals for treatment effect estimates Assessment of confounding, statistical attempts to correct for imbalance Assessment of heterogeneity, quality assurance, stratification
Results	Treatment effect estimates with confidence intervals, P-value for primary hypothesis
Discussion	Conclusions supported by results with possible biases and limitations taken into consideration
Funding or sponsorship	Type and sources of funding, potential conflicts of interest

[a] Adapted from Refs. [6,10,16].

Table 2

Example of a simple study quality instrument: the Jadad scale

Was the study described as randomised?	0/1
Was the randomisation method described and appropriate (table of random numbers, computer-generated, etc.)?	0/1
Was the study described as double blind?	0/1
Was the method of double blinding described and appropriate (identical placebo, active placebo, dummy, etc.)?	0/1
Was there a description of withdrawals and dropouts?	0/1
Deduct one point if the randomisation method was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc.).	0/ − 1
Deduct one point if the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy).	0/ − 1

of all, although its simplicity is an attractive feature of the Jadad scale, it is clear that it only addresses a limited component of study quality. It is easy to imagine a trial scoring 5 on the Jadad scale but still failing to reach an acceptable quality in other aspects. Secondly, while it is quite reasonable to believe that blinding is very important in pain assessment, it may be of much less importance, or even impossible to achieve, in many other clinical situations.

The use of quality scores in systematic reviews and meta-analysis is quite controversial and raises a number of interpretational and statistical issues that are beyond the scope of the present paper. The simplest approach is to define a threshold quality score for inclusion in the analysis. However, this is somewhat arbitrary and attempts have been made to devise a quality factor that can be used as a weight in a pooled analysis of treatment effect. The interested reader may find a good entry point to the discussion of theses issues in the report by West et al. [6] and the references therein.

Grading based on study design

The most widely used classification systems for strength of medical evidence are based on study

Table 3

Simplified scale for classification of evidence based on underlying study design[a]

Level I	Adequately powered, high quality randomised trial, or meta-analysis of randomised trials showing statistically consistent results
Level II	Randomised trials inadequately powered, possibly biased, or showing statistically inconsistent results
Level III	Non-randomised studies with concurrent controls
Level IV	Non-randomised studies with historical controls (i.e. typical single arm phase II studies)
Level V	Expert committee review, case reports, retrospective studies

[a] Resulting from discussion between Buyse, Bentzen, Tannock and Therasse (May 2003).

design. Unfortunately, also in this case there is no consensus on a single preferable system but most systems are variations on the theme set out in Table 3. At level I there is only evidence originating from adequately powered randomised controlled trials, either individual trials or a meta-analysis of trials. While the higher priority given to randomised controlled trials is fully justified, it is easy to imagine that there are poor randomised trials that may be less convincing than well-designed studies at lower levels of evidence. Some groups have modified and expanded this system. For example, the Oxford Centre for Evidence-based Medicine [9] has suggested a modification to indicate whether the available evidence at a given level is heterogeneous or not. While there is a lot of merit to this proposal, it is not widely used and there is of course a general problem with these systems being so elaborate that few people will remember them by heart.

Quality of reporting: the CONSORT guidelines

Arguably, the most ambitious and in many ways successful attempt to help improve the quality of reports on RCTs is the CONSORT (CONsolidated Standards Of Reporting Trials) statement [10] that was developed by a group of statisticians and biomedical journal editors and published in 1996 with a subsequent minor update in 2001 [11]. The statement included a checklist of items related to the design, conduct and analysis of a trial, that the group suggested should be reported as a minimum in a publication of the outcome of the trial. Many of these items had previously been shown empirically to be associated with the magnitude of effect estimates in various trials. Since 1996, more than 130 journals have adopted the CONSORT statement as a minimum requirement for trial design and analysis aspects that should be included as a minimum in reports on RCTs.

Moher et al. [12] compared reports on RCTs in 1994 and 1998, i.e. before and after the CONSORT statement was published, in three journals (*British Medical Journal*, *JAMA* and *The Lancet*) that did, and in one journal (*New England Journal of Medicine*) that did not adopt these guidelines. The authors concluded that adoption of the CONSORT statement had improved the quality of reporting of trials. However, it should be noted that this conclusion was based on an increase in the number of items on the CONSORT checklist that were reported in papers in the journal or quality assessed on scales having a considerable overlap with the CONSORT list.

At the conclusion of this section, it should be stressed once again, that the CONSORT statement only relates to the completeness of reporting. This is not a surrogate for study quality as illustrated in the paper by Huwiler-Muntener et al. [13]. Clearly, the usefulness of the published medical literature on RCTs will be improved with a more detailed and more standardised reporting. However, compliance to the CONSORT statement should be regarded a necessary but clearly not sufficient criterion for acceptable standards of reporting and it is not in itself a measure of study quality.

Generalisation of study findings

A controlled clinical trial is a model of standard clinical practice that is set up to eliminate the possibility of systematic bias. At the same time, most protocols attempt to standardise as much as possible various aspects of diagnosis, therapy and follow-up, and to define inclusion and exclusion criteria that will reduce the heterogeneity of biological characteristics in the trial population. The underlying hypothesis is that the results of the trial can be applied to future patients treated in routine clinical practice. However, a number of studies suggest that this may not always be the case. One example is the study by Feuer et al. [14] who compared patients with advanced testicular cancer from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) registry programme with diagnostically comparable concurrent trial patients from the Memorial Sloan–Kettering Cancer Center (MSKCC). A number of exclusion criteria were applied in both series to make them as comparable as possible at baseline. Nevertheless, MSKCC patients had a significantly better survival than SEER patients in both the minimal/moderate extent of disease and the advanced disease categories.

The authors found no obvious explanation for this difference but speculated that this might be explained by the trial patients receiving better care.

Study population

New therapies are often tried in fairly restrictively defined populations. Inclusion criteria frequently include performance status or stage of disease. Exclusion criteria could for instance typically comprise various types of co-morbidity or lab test values outside a defined range. While this makes good sense from a biological and statistical point of view, it may affect the generalisability of the therapeutic outcome. Due to the random treatment allocation, the comparison of two therapies will, on average, be unbiased. However, it is biologically and clinically conceivable, that a therapeutic gain in a highly selected population of prognostically favourable cases with no major comorbidity will to some extent be diluted or even be completely lost in a population of unselected cases.

Diagnostic intensity

Look carefully at the diagnostic work-up of patients before entering the trial. While proper randomisation on average ensures an unbiased comparison of treatment effect, it is conceivable that the relative benefit of an intervention would depend on burden and/or site of disease. Assume that a drug shows a 30% complete response rate in patients with metastatic disease. Clearly, this response rate will be expected to be lower in a population of patients with metastatic disease identified by a less sensitive diagnostic programme. Likewise, the response of metastases in different sites may vary considerably. Another example would be a trial accruing patients with M_0 disease. Here it is possible that by aggressive diagnostic screening a higher proportion of patients with a metastatic phenotype is removed from the trial population, and this again may introduce a biological selection of trial cases.

Interventions

Details of interventions are often under-reported in the literature. Any kind of supportive care, specific diagnostic procedures to detect emerging toxicity, interventions in case of toxicity and eventual salvage therapy in case of progression are all part of the package deal tested in the trial. All of these should be described in sufficient detail to allow a reader to evaluate and eventually to reproduce the therapeutic strategy. It is important to establish whether the same level of supportive care was provided in the conventional arm of the trial. It has also been convincingly argued that in trials comparing active therapy with best supportive care, the latter is also an intervention that should be described in sufficient detail to allow it to be reproduced.

Length and intensity of follow-up

The benefit of one therapy over another may change with time. An example is provided by Bartelink and colleagues [15] who used a factorial 2 × 2 design in an RCT of radiotherapy alone or combined with chemotherapy or/and hormonal therapy in 410 patients with locally advanced breast cancer. Four analyses were performed: during the trial, immediately after closure and 8 and 13 years after the start of the trial. Chemotherapy showed a statistically significant advantage at the two early analyses but not at longer follow-up; for hormonal therapy it was the opposite: a statistically significant benefit being observed at the time of the last analysis but not at the earlier analyses. Thus, an early outcome favouring the chemotherapy arm was subsequently changed into a relatively more pronounced benefit from hormonal therapy. Studies with relatively short follow-up may show initial differences that may not hold-up as the trial data mature. The relevant timescale for judging efficacy depends not just on the mechanism of action of the involved therapies but also on the natural history and the life-expectancy of the patient. Differences in follow-up time especially between studies should be considered as a possible confounder in a direct comparison of efficacy.

Endpoints

Endpoints are the health-related events, which may or may not occur in an individual patient, that are used in quantifying the relative success of a therapy. Examples are overall survival, local tumour control, time to progression, toxicity or quality of life. Endpoints should reflect the treatment aim. If therapy is aimed at palliation, the endpoint should reflect symptom relief. In this case, survival or tumour control endpoints will only be of secondary importance. Unfortunately, there is no consensus on the exact definition of various endpoints. As an example, when a paper refers to metastasis-free survival there are two possible definitions of what constitutes an "event". In both cases, the occurrence of a clinically manifest metastasis will be considered a treatment failure. The difference arises when considering patients who die before they develop a metastasis. In one case, death is considered an event; in the other case, death is treated as a cause of censoring. It has been suggested [16] to use "metastasis-free survival" to denote the former of these definitions, i.e. a patient will have to be alive and free from distant metastasis to count as a success. The latter definition is then referred to as the "metastasis-free rate" as a function of time.

In case of disease-related endpoints, infrequent follow-up, the lack of specific diagnostic procedures at follow-up and a low autopsy frequency will all tend to decrease not only the rate of disease progression but also the ability to resolve a change when comparing therapies.

For the critical reader, the exact definition of endpoints is of major importance in evaluating the findings of a study. Regrettably, even in leading cancer journals, definitions of endpoints are often omitted from the Materials and Methods section of research papers. Altman et al. [17] surveyed 132 papers reporting survival analyses published in the *British Journal of Cancer, European Journal of Cancer, Clinical Oncology* and *American Journal of Clinical Oncology* between October and December 1992. One or more endpoints were not clearly defined in 62% of the papers. A particular problem turned out to be the definition of time to progression where it was unclear in 39 of 64 papers (61%) using this endpoint how death was included in the analysis.

Actuarial statistics

Many endpoints in cancer trials require prolonged follow-up of the patient. Special statistical methods, often referred to as actuarial statistics or survival statistics, are required to allow for incomplete or censored observations, that is, patients who had not reached the endpoint at the time of the last follow-up. Several statistical tests are available for comparing two time-of-occurrence endpoints. The most widely used is the logrank test, sometimes referred to as the Mantel–Cox test. An alternative is the Gehan–Breslow test and the main difference between the two is that the latter test gives relatively more weight to short observation times. A brief introduction to these methods has been given by Buyse [18]. Time-censored data that are analysed without using appropriate actuarial methods may be misleading.

Toxicity and treatment-related morbidity

Toxicity is an important aspect of any active cancer therapy. Unfortunately, published trials often under-report early and late toxicity and comparisons between studies are hampered by the lack of a uniform system for grading and reporting toxicity. The US National Cancer Institute's Cancer Therapy Evaluation Programme has supported the development of the Common Toxicity Criteria version 2.0 [19] which is a dictionary that can be used for grading early toxicity after radiotherapy, drug therapy or surgery. A revision has just been published, the Common Terminology Criteria for Adverse Events version 3.0, and this dictionary define terms and grades for late effects of cancer therapy as well. A number of methodological issues in the reporting of early and late effects of cancer therapy have been discussed recently [20].

Quality of life

Quality of life endpoints play an important role especially in palliative studies. Proper analysis of this kind of data poses several problems both from a statistical as well as an interpretational point of view which it is beyond the scope of the present paper to discuss. For a recent review see Ref. [21].

Therapeutic gain

One last issue to be briefly mentioned here is the lack of a consensus on how best to quantify therapeutic gain. This is of course particularly of interest if a specific treatment both improves tumour outcome and is associated with increased toxicity. At the moment, there is no generally accepted way of quantifying this trade-off and this is definitely an area where more research is needed.

Summarising the outcome of a trial

There are a number of principles and pitfalls in the interpretation of trial results that the critical user should be aware of.

The intention-to-treat principle

The primary analysis of an RCT should be performed according to the intention-to-treat principle and this means that all randomised patients are included in the analysis in the treatment arm to which they were randomly allocated. This is irrespective of any deviations from the allocated therapy, non-compliance to the prescribed treatment and any violation of inclusion or exclusion criteria. Any patient exclusion after the time of randomisation should be treated with suspicion as they may very well introduce a bias in the effect estimates. At the same time, non-adherence to this principle may to some extent be a surrogate for poor trial quality.

P-values or confidence intervals

In 1925, the Cambridge geneticist Ronald A. Fisher published his textbook *Statistical Methods for Research Workers* which became one of the most influential texts ever published on biomedical research methodology. In the book, Fisher laid down the foundations of classical hypothesis testing and it was here Fisher proposed that if a difference observed between two groups had a less than 5% probability of occurring by chance, this difference was "significant". This convention remains in use today, almost 80 years after it was first proposed, despite it being completely arbitrary.

Medical research in particular is obsessed with significance tests or "P-values". To a large extent these have become synonymous with "real" research and "significance" is often directly equated with importance. In reality, much more information is provided by an estimate of the difference in therapeutic effect with 95% confidence intervals (C.I.). A useful interpretation of the 95% C.I. is that this is the set of hypotheses that would still be consistent with (i.e. not significantly different from at the 5% level) the result of the trial.

Fig. 1 shows treatment effects with 95% C.I. for five hypothetical trials. The first three of these all have a non-significant P-value, in other words a zero difference is included in the 95% C.I. Still, the information content in the trial with 2 x 600 patients is clearly much higher than in the two other cases: the confidence interval is sufficiently narrow to exclude any major benefit from one treatment relative to the other. The two last "trials" are both statistically significant. Nevertheless, the trial with 2 x 30 patients tells us very little about the magnitude of this therapeutic gain as the "set of acceptable hypotheses" range all the way from a 2% to a 51% benefit.

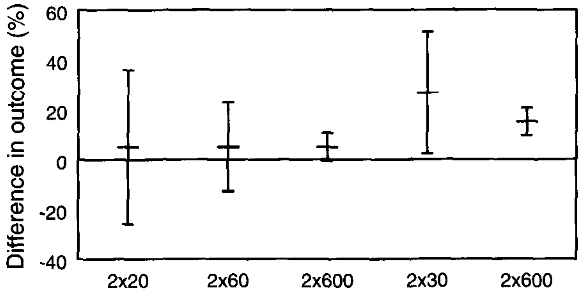

Fig. 1. Confidence intervals for the difference between two proportions estimated from a binomial distribution for various sample sizes in the two groups.

"A difference is only a difference if it makes a difference"

A statistically significant difference between two treatments may be so small that it is clinically irrelevant. This is yet another illustration of how much more information is presented in an effect estimate with 95% C.I.

Reading and understanding a report on a 'negative' trial

Labelling trials as "positive" and "negative", depending on whether they report a statistically significant difference between treatment arms, is of course biased language. A trial finding no significant difference between alternative therapies may contribute very valuable knowledge and have a positive impact on cancer patient care. Nonetheless, it reflects the culture of clinical research: each time a novel therapy is tried, we all hope that it represents a therapeutic benefit.

"Absence of proof is not proof of absence"

Look at the confidence intervals! It is as simple as that — provided of course that the paper actually reports 95% C.I. As discussed above, the 95% C.I. is the set of acceptable hypothesis that would still be consistent with the outcome of the trial. The fact that we cannot detect a difference does not mean that it is not there. It simply means that the therapeutic benefit is most likely too small to be detected in a trial of this size. Increasing the statistical power, by increasing the sample size, will narrow the confidence interval and this means that we can rule out successively smaller benefits by conducting larger and larger trials (see Fig. 1). Alternatively, as discussed below, the outcome of several trials addressing the same question may be pooled in a meta-analysis.

The precise interpretation of a non-significant trial is of particular importance if the trial tests a biological principle. Again, a non-significant difference does not show that the underlying effect does not exist but only that its influence on outcome is likely to be below some upper bound, typically given by the upper 95% C.I.

The number of patients, N, required in a trial de-

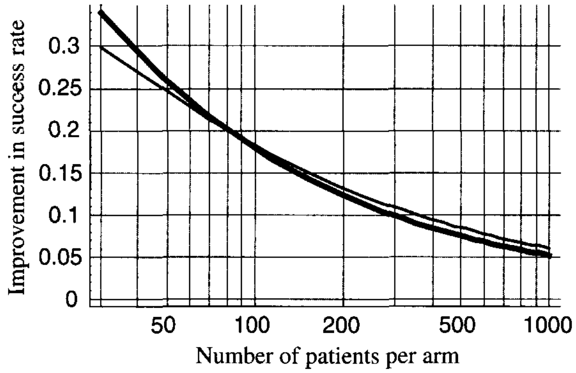

Fig. 2. Improvement in success-rate required in order to be significant at the 5% level with an 80% probability (=power) as a function of the number of patients per arm in a two-arm trial. The relationship is plotted for a baseline success-rate of 0.6 (thin line) or 0.2 (thick line).

pends on the level of the test, conventionally chosen as $\alpha = 5\%$, the power, $1 - \beta$, or statistical resolution of the trial, often chosen as 80% or 90%, and the treatment effect, say a difference, Δ, in 5-year survival, that the trial is designed to resolve. If we specify three of these four quantities, N, α, β and Δ, the fourth of them can be calculated. Fig. 2 shows the treatment effect, that a trial with $\alpha = 5\%$ and a power of 80% can resolve as a function of the number of patients per treatment arm. From the graph it can be seen that a trial with 2 x 300 patients can roughly detect a 10 percentage point increase in survival with a 5% level of the test and a power of 80%. A trial with 2 x 1000 patients will allow resolution of a 5 percentage point improvement. At the other end of the scale, a trial with 2 x 80 patients randomised can resolve a change from 20% to 40% survival, which would be nothing short of a miracle in clinical oncology. This figure may provide a quick impression of the statistical power of a trial where a power calculation is not included in the report.

Many published trials are underpowered to detect the magnitude of a treatment effect that can reasonably be expected. A review [16] of 92 radiotherapy RCTs showed that the median sample size was 2 x 72 patients. As seen from the graph, a trial of this size would need a 21 percentage point improvement in success-rate to achieve 80% power.

Factors affecting the resolution of a trial

There are a number of trial characteristics — in addition to sample size — that may affect the ability of a trial to detect a treatment benefit of a given magnitude. A heterogeneous patient population, patient-to-patient variability in treatment characteristics, diagnostic and follow-up procedures, variability among centres in patient population or quality of care, lack of appropriate quality assurance, short follow-up, high rate of loss of patients to follow-up are all factors that will reduce the ability of a trial to detect a real difference between treatment arms.

Equivalence trials

A special case of non-significant trials is the equivalence trial, a trial designed from the onset to demonstrate that two therapies give an equivalent outcome. In a certain sense, a null result is a positive finding of such a trial. For the critical reader, this means that any factor reducing trial resolution, including the factors listed in the previous section, would most likely bias the outcome in favour of the investigators' prior beliefs. Or to put it even more bluntly: a poorly designed and conducted trial will maximise the chance of demonstrating equivalence!

Reading and understanding a report on a "positive" trial

A number of issues should be considered by the critical user when reading a report of a trial claiming a statistically significant advantage of one therapy over another.

Multiple comparisons

At the core of classical hypothesis testing is the idea that the probability of a Type 1, or false-positive, error should be 5% or lower. Testing at the 5% level means exactly that: even if in reality there is no difference between the two groups we compare, 1 in 20 of the tests we perform will result in a "statistically significant difference". Clearly, with an increasing number of independent tests the chance will increase rapidly that at least one of them would come out significantly just by chance. If we restrain ourselves to testing a single hypothesis, this 5% level may be reasonable, but in reality most reports on clinical trials contains dozens of P-values. Pocock et al. [22] looked at the number of significance tests of differences in treatment effect in 45 reports on clinical trials published in the *British Medical Journal, The Lancet*, and the *New England Journal of Medicine*. The median number was eight and these were typically tests of multiple endpoints, subgroup analyses or repeated looks over time. Tannock [23] conducted a similar review but in addition to the actual number of P-values reported he tried to estimate the likely

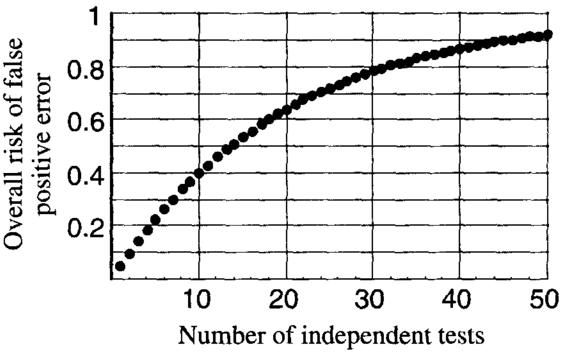

Fig. 3. Overall probability that at least one of a number of independent hypothesis tests will come out significant at the 5% level, even if there is no real difference between the groups compared in any of the tests.

number of tests that had been performed but not necessarily reported. The median number of significance tests for major outcome parameters was six, whereas the median number of reported and implied tests were 13. Including also subgroup analyses the median number of efficacy tests was 20.

Fig. 3 shows the probability of finding at least one significant P-value at the 5% level as a function of the number of independent statistical tests performed. This figure allows a simple impression of the effective significance level of a study applying multiple testing. In case of 13 significance tests, the chance of at least one of them coming out significantly is nearly fifty-fifty, 48% to be precise. With a median number of efficacy tests of 20, the chance is nearly 2:3. If should be noted that this calculation is conservative in case of non-independence of the tests. Nevertheless, it does produce a useful impression of just how significant a $P < 0.05$ is in studies with multiple tests. There are statistical corrections for multiple testing; the most frequently used is probably the Bonferroni correction. One may ask why such corrections are not always used. There are, however, situations where this would not be reasonable. For a slightly more detailed discussion see Ref. [16]. The most sensible remedy against multiple testing is to define a primary endpoint and a corresponding primary hypothesis *before* a trial is started.

The wide availability of inexpensive computers and statistical software packages has undoubtedly increased the problem with multiple comparisons: a review found that the proportion of papers in *Arthritis and Rheumatism*, comparing two groups on more than 10 variables at the 5% level, rose from 6% in 1967–1968 to 38% in 1982.

Subset analysis

As mentioned above, a common cause of multiple comparisons is the systematic testing of differences between treatment arms in various subgroups. In the review by Pocock et al. [22], 23 out of 45 trials (51%) reported at least one subgroup analysis of treatment effect. None of these papers specified that the subgroup analysis had been planned prospectively. P-values for treatment effects in subgroups should mainly be seen as measures of association rather than be interpreted in the standard framework of hypothesis tests. Thus, subgroup analysis should mainly be regarded as a hypothesis-generating exercise — the hypotheses arising from this should be tested subsequently in independent studies.

Early/late stopping, trial monitoring

All prospective clinical trials should have a defined target sample size and this should be justified by clinical or biological arguments regarding the expected therapeutic difference the trial should be powered to detect. When the trial is analysed and published it is important to look for a deviation between the actual and the planned sample size. If a trial is stopped early, or if it continues accrual after the target sample size is reached, this decision should be justified in the paper and should be made without looking at whether or not there is a significant difference between treatment arms. The decision to stop or continue a trial should not be made by the trialists but rather by an Independent Data Monitoring and Ethics committee. Multiple looks and a decision to stop or continue a trial based on the significance of an emerging difference will lead to a biased estimate of the P-value.

Publication bias

Trials showing a statistically significant difference between the outcomes of the various treatments are more likely to be published in oral and written form than trials showing no significant difference. This so-called publication bias leads to a bias in the evaluation of the effectiveness of therapies reported in the literature. Easterbrook et al. [24] looked at the publication status of 487 research projects approved by the Central Oxford research Ethics Committee between 1984 and 1987. At the time of Easterbrook's review, data from 285 of the projects had been analysed and results from 52% of these had been published. Studies finding a statistically significant difference between groups had a significantly higher chance of having been published and this was also the case in a multivariate analysis including co-variates such as study design, sample size and source of funding.

Krzyzanowska et al. [25] reviewed all large randomised controlled trials (defined as trials with a sample size of more than 200) presented at annual meetings of the American Society of Clinical Oncology (ASCO) between 1989 and 1998. Using the Kaplan–Meier method to correct for censoring, the authors found that 74% of the 510 identified trials had been published in written form within 5 years after presentation at the meeting. Publication status was estimated according to the statistical significance reported in the meeting abstract: 81% of the studies with significant results had been published by 5 years compared to 68% of the studies with non-significant results ($P = 0.0001$). The 10-year publication rate estimates were more than 90% for the significant versus 80% for the non-significant abstracts.

This study documents that even in a sample of moderate-to-large sized RCTs, there is still a bias against written publication of non-significant studies. It could be speculated that submission of an abstract to the prestigious ASCO (American Society of Clinical Oncology) meeting is likely to create a selection bias favouring subsequent publication of the study findings as abstracts from non-significant studies could be biased towards being more informative (e.g. originate from trials set up to demonstrate equivalence of two interventions) or originate from groups or individuals with a stronger-than-average push towards publishing their research.

A special form of publication bias arises from duplicate publication of data from trials finding large treatment effects. Tramer et al. [26] surveyed 84 RCTs of ondansetron's effect on post-operative emesis published between 1991 and 1996 and found that data from nine trials had been published in 14 further papers, and none of these had a clear reference to any prior publication. Studies were more likely to be published repeatedly if they showed a large treatment effect and a meta-analysis including duplicated data gave a 23% overestimation of the efficacy of ondansetron compared with an analysis without duplicated data.

Systematic overviews and meta-analysis

Review papers are an important source of information in evidence-based medicine. Traditionally, the narrative review has been the most common type (Table 4). Often authored by a leading expert in the field and with some scope for personal beliefs and polemics, these reviews may be both entertaining and enlightening. However, they do require critical reading from the user's side and perhaps they work best when they are part of an ongoing discussion between informed debaters.

Many researchers have felt that the traditional narrative review was of little value in establishing an evidence-based best practice and this has been one of the driving forces in the evolution of the systematic overview. The aim was to develop a more objective type of overview with a quantitative synthesis, a so-called meta-analysis, of information from many independent trials into an overall best estimate of therapeutic effect (Table 4).

Meta-analysis

From a statistical point of view, meta-analysis is a fairly straightforward or even trivial method. Basically this is just an example of a stratified analysis, a well-established technique described in any standard textbook on statistics. It is the systematic and often uncritical application of this technique in various fields of medicine that has made meta-analysis a hot and even contentious issue in evidence-based medicine.

A tutorial review of the methods of and the case for meta-analysis has been published recently by Pignon and Hill [27]. Arguably, meta-analysis is useful as an element in a systematic overview. However, also in this case the critical user should be aware of a number of fundamental assumptions and limitations that may affect the interpretation of even the largest and most professionally conducted meta-analysis.

Individual patient vs. summary trial outcome data

Publication bias is a major concern with meta-analyses. Simes [28] compared the difference in estimated therapeutic effect of an alkylating agent versus combination chemotherapy in advanced ovarian cancer. The pooled results of 16 published trials showed a significant advantage of combination chemotherapy. However, the pooled result of 13 registered but unpublished trials showed no significant difference between the two types of therapy.

Stewart and Parmar [29] conducted a meta-analysis of non-platinum single drug versus platinum-based combination chemotherapy in advanced ovarian cancer. Pooling the published effect estimates from 8 trials comprising 788 patients showed a 7.5% absolute survival advantage at 30 months after combination chemotherapy ($P = 0.027$). However, analysing the individual data from 1329 patients in 11 trials, includ-

Table 4
Three archetypical classes of reviews found in the literature

Domain	Traditional narrative review	Traditional meta-analysis	Systematic narrative review
Study inclusion	Unclear inclusion criteria	Search strategy stated in paper	Search strategy stated in paper
	Often plagued by selective citation of references/exclusion of studies	All studies included	Motivated exclusion of studies
	Non-randomised studies and case reports often overly emphasised if they support the reviewer's point	Randomised trials only	Give priority to randomised trials; observational studies discussed when relevant with an assessment of possible biases and confounding factors
	Abstracts and personal communications may be included	Includes "grey" literature, published abstracts, unpublished data	Give relatively more weight to published and peer-reviewed reports
Heterogeneity	Emphasise differences between studies; attribute differences in outcome to details of therapy	De-emphasise differences between studies; broadening the question to get larger numbers	Systematic analysis of heterogeneity
Variability in trial outcome	Clinical, biological and treatment related factors are assumed to be the main sources of variability	Sampling variability is assumed to be the main source of variability in outcome	Explore whether variability in outcome among studies is informative
Main outcome	A motivated (subjective?) recommendation	A pooled P-value and a pooled estimate of treatment effect	A motivated (objective?) recommendation
More trials?	In danger of producing a false impression of the current level of knowledge	Discourages more trials as very large trials will be required to change the outcome of a cumulative meta-analysis	Encourages and suggests trials addressing specific outstanding issues
Meta-analysis?	"A toy for statisticians..." [accompanied by shrug of shoulder]	"The best thing since sliced bread. . . "	"May be useful as one element in a critical, systematic review"
"Typical" conclusion	"A randomised trial would not be feasible/ethical/necessary. . ."	"A small, but highly significant effect. . . "	"Current best evidence suggests. . . " "Recommended directions for future research include. . ."

ing the 8 published trials, showed a non-significant survival advantage of combination chemotherapy of just 2.5% at 30 months. The authors explained the discrepancy between the two analyses as the result of publication bias, patient exclusions and shorter follow-up in the published reports that all tended to overestimate the advantage of combination therapy.

Tests for heterogeneity

A recent meta-analysis of chemotherapy combined with loco-regional therapy in squamous cell carcinoma of the head and neck [30] showed a highly significant survival benefit when the modalities were administered concurrently. The absolute gain in 5-year survival was estimated at 8%. However, the test for heterogeneity was also highly significant, $P < 0.0001$. This means that the basic assumption of meta-analysis is violated, the variation in trial outcome cannot be assumed to be due to sample size alone. The inevitable conclusion is that some reports have presented major gains from concurrent therapy, whereas others have not. The 8% estimated gain in survival from pooling these studies becomes very difficult to interpret in this situation: why would one be interested in the average gain from some strategies that work and others that do not?

In many other meta-analyses the test for heterogeneity is not significant. However, this is arguably a low-powered test in many situations. In case of an ordinal or continuous variable, a test for trend would be more powerful than the standard heterogeneity test.

The issue of heterogeneity is very important, not just for the interpretation of the meta-analysis in itself, but also because there is potentially useful knowledge that may be derived from heterogeneities in therapeutic effect.

The systematic narrative review

Perhaps the most worrying criticism of meta-analysis has come from a consideration of their external, rather than internal, validity. LeLorier et al. [31] looked at the ability of meta-analyses to predict the outcome of large randomised controlled trial addressing the same question. Their material consisted of 12 randomised controlled trials each with a sample size of more than 1000 patients and 19 meta-analyses on the same question as one of the trials. They defined a "negative" outcome of a meta-analysis or a trial as no statistically significant benefit from the intervention studied, and a "positive" outcome as a statistically significant difference. The negative predictive value of the meta-analyses was 67% and the positive predictive value was 68%.

Another recent example is the contrasting conclusions regarding the survival advantage of post-operative radiotherapy for breast cancer where three randomised trials [32–34] comprising more than 3,400 patients consistently showed a survival advantage in contrast to the lack of any significant difference seen in a large meta-analysis including more than 17,000 patients [35]. An interesting analysis by van de Steene et al. [36] showed a clear trend over time towards increasing benefit from post-operative radiotherapy in more recent trials. There are good reasons to believe that modern radiotherapy techniques can largely avoid the increased mortality from ischaemic heart disease in patients treated with post-mastectomy radiotherapy in earlier series [37].

It is also instructive to consider the size of a trial required to substantially change the outcome of the meta-analysis. Assume that the real benefit from modern radiotherapy is the increase in survival from 45% to 54% observed in the Danish trial [32]. Assume further that it would take at least a 5% improvement in survival to convince the doubters that radiotherapy is indicated in high-risk patients. Under these assumptions, the Danish trial should have accrued 7680 patients to outweigh the accumulated "evidence" in the meta-analysis. This is nearly five times the number of patients actually accrued in the trial among pre-menopausal women, and it is clear that if this had been the basis for deciding whether to embark on the trial, the trial would never have been started. In this sense, a meta-analysis may discourage further trials as it will appear to provide overwhelming evidence for or against a specific therapy that will not substantially change even if a large new RCT were to come out with a different result.

These examples illustrate the potential negative side effects of a blind reliance on meta-analyses. The ideal would be a systematic narrative review, combining the rigour and transparency of classical meta-analysis in the identification of relevant data with the afterthought and critical analysis of the classical expert review. Table 4 lists some of the desirable characteristics of a systematic narrative review. Admittedly, this author can think of only a few reviews published to date that qualify for this label!

The sociology of clinical trials

Research is ideally an objective process but in reality there are a number of factors that may potentially affect the credibility of research.

Journal impact factors and target groups

There is, to the best knowledge of this author, no evidence whatsoever that the quality of trials published in high-impact factor journals is higher than average. Nevertheless, there is a tendency that such trials are assigned more credibility in daily discussions. For the critical user this is yet another trap to be avoided. Even the most distinguished journals occasionally publish poor quality science and the most informative trial on a given topic may well be published in a rather obscure journal.

Another issue is language bias: MEDLINE has an under-representation of non-English papers and this is also true for many systematic reviews. Studies have shown that there is no significant difference between the quality of trial reporting in English and the quality of trial reporting in other major languages [38].

Research ethics

All human experiments, including clinical trials, are regulated by international codes of conduct that are almost uniformly accepted worldwide. Nevertheless, there are studies that are borderline in the sense that they are designed or conducted in a way that would be unacceptable to research ethics committees in some countries. An example is pre-randomisation according to Zelen's principle, which twenty years ago was considered ethically acceptable but is not anymore in several countries. The critical user should consider that the outcome of such trials may not be easily reproduced in other countries and therefore should probably be assigned relatively less weight or in extreme cases be disregarded.

Sponsorship and conflicts of interest

The cost of bringing a new drug on the market in the USA has been estimated at US $500 million [39]. Clearly, an investment of this magnitude creates a potential conflict of interest for anyone employed by or having financial interests in the company developing a new drug. Perhaps more worrying is the fact that a Cochrane review of drugs used in treatment of schizophrenia showed a larger effect estimate in company-sponsored trials compared to trials without obvious sponsorship [40]. Many journals have a policy that conflict of interests should be declared in the paper, but this is not always enforced. Evidence from sponsored trials can of course not be disregarded. However, sponsorship may be taken into account when evaluating a body of perhaps conflicting evidence on a given therapy.

Scientific misconduct and fraud

Fabrication of data is extremely rare and although there have been a couple of high-profile cases in clinical cancer research [41,42], this is an issue that can generally be put aside for the critical user. A set of guidelines on publication ethics have been developed and it may be instructive to read these as a way of increasing the awareness of scientific misconduct and related issues [43].

Towards evidence-based oncology

Evidence-based medicine is a goal, a dynamic state that can only be approached asymptotically. Arguably, the number of high-quality randomised controlled trials is increasing. However, much of the published literature can only be labelled as poor quality. Users of medical evidence, i.e. primarily treating physicians and researchers, will have to develop their critical appraisal skills as a bulwark against misleading or exaggerated claims in the published literature.

References

1 Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996, 312: 71–72.

2 Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. JAMA 1998, 280: 237–240.

3 Meric F, Bernstam EV, Mirza NQ, et al. Breast cancer on the world wide web: cross sectional survey of quality of information and popularity of websites. BMJ 2002, 324: 577–581.

4 Bailey KR. Generalizing the results of randomized clinical trials. Control Clin Trials 1994, 15: 15–23.

5 Evans JT, Nadjari HI, Burchell SA. Quotational and reference accuracy in surgical journals. A continuing peer review problem. JAMA 1990, 263: 1353–1354.

6 West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No. 02-E016. 2002. Rockville, MD, Agency for Healthcare Research and Quality.

7 Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999, 282: 1054–1060.

8 Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996, 17, 1-12,

9 Oxford Centre for Evidence Based Medicine. Levels of evidence and grades of recommendation. Accessed 28 Apr. 2003. Web page. Available at *www.cebm.net/levels_of_evidence.asp*

10 Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. JAMA 1996, 276: 637–639.

11 Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001, 285: 1987–1991.

12 Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001, 285: 1992–1995.

13 Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA 2002, 287: 2801–2804.

14 Feuer EJ, Frey CM, Brawley OW, et al. After a treatment breakthrough: a comparison of trial and population-based data for advanced testicular cancer. J Clin Oncol 1994, 12: 368–377.

15 Bartelink H, Rubens RD, van der Schueren E, Sylvester R. Hormonal therapy prolongs survival in irradiated locally advanced breast cancer: A European Organization for Research and Treatment of Cancer randomized Phase III trial. J Clin Oncol 1997, 15: 207–215.

16 Bentzen SM. Towards evidence based radiation oncology: Improving the design, analysis, and reporting of clinical outcome studies in radiotherapy. Radiother Oncol 1998, 46: 5–18.

17 Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer 1995, 72: 511–518.

18 Buyse ME. Clinical trial methodology. In: Peckham M, Pinedo HM, Veronesi U, editors. Oxford Textbook of Oncology. Oxford: Oxford University Press, 1995: 2377–2395.

19 Trotti A, Byhardt R, Stetz J, et al. Common toxicity criteria: version 2.0. an improved reference for grading the acute effects of cancer treatment: impact on radiotherapy. Int J Radiat Oncol Biol Phys 2000, 47: 13–47.

20 Bentzen SM, Doerr W, Anscher MS, Denham JW, Hauer-Jensen M, Marks LB, et al. Normal tissue effects: reporting and analysis. Sem Rad Oncol. In press.

21 Cella D, Chang CH, Lai JS, Webster K. Advances in quality of life measurements in oncology patients. Semin Oncol 2002, 29: 60–68.

22 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987, 317: 426–432.

23 Tannock IF. False-positive results in clinical trials: Multiple significance tests and the problem of unreported comparisons. J Natl Cancer Inst 1996, 88: 206–207.

24 Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991, 337: 867–872.

25 Krzyzanowska MK, Pintilie M, Tannock IF. Failure to publish large randomized trials presented at an oncology meeting. JAMA. In press.

26 Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ 1997, 315: 635–640.

27 Pignon JP, Hill C. Meta-analyses of randomised clinical trials in oncology. Lancet Oncol 2001, 2: 475–482.

28 Simes RJ. Confronting publication bias: a cohort design for meta-analysis. Stat Med 1987, 6: 11–29.

29 Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993, 341: 418–422.

30 Pignon JP, Bourhis J, Domenge C, Designe L. Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. MACH–NC Collaborative Group. Meta-Analysis of Chemotherapy on Head and Neck Cancer. Lancet 2000, 355: 949–955.

31 LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997, 337: 536–542.

32 Overgaard M, Hansen PS, Overgaard J, et al. Postoperative radiotherapy in high-risk premenopausal women with breast cancer who receive adjuvant chemotherapy. N Engl J Med 1997, 337: 949–955.

33 Overgaard M, Jensen MB, Overgaard J, et al. Postoperative radiotherapy in high-risk postmenopausal breast-cancer patients given adjuvant tamoxifen: Danish Breast Cancer Cooperative Group DBCG 82c randomised trial. Lancet 1999, 353: 1641–1648.

34 Ragaz J, Jackson SM, Le N, et al. Adjuvant radiotherapy and chemotherapy in node-positive premenopausal women with breast cancer. N Engl J Med 1997, 337: 956–962.

35 Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery in early breast cancer. An overview of the randomized trials. N Engl J Med 1995, 333: 1444–1455.

36 Van de Steene J, Soete G, Storme G. Adjuvant radiotherapy for breast cancer significantly improves overall survival: the missing link. Radiother Oncol 2000, 55: 263–272.

37 Hojris I, Overgaard M, Christensen JJ, Overgaard J. Morbidity and mortality of ischaemic heart disease in high-risk breast-cancer patients after adjuvant postmastectomy systemic treatment with or without radiotherapy: analysis of DBCG 82b and 82c randomised trials. Radiotherapy Committee of the Danish Breast Cancer Cooperative Group. Lancet 1999, 354: 1425–1430.

38 Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews . Lancet 1996, 347: 363–366.

39 Davidoff F, DeAngelis CD, Drazen JM, et al. Sponsorship, authorship, and accountability. N Engl J Med 2001, 345: 825–826.

40 Wahlbeck K, Adams C. Beyond conflict of interest. Sponsored drug trials show more-favourable outcomes. BMJ 1999, 318: 465.

41 Jenks S. Congressional hearing delves into NSABP fraud issues. National Surgical Adjuvant Breast and Bowel Project. J Natl Cancer Inst 1994, 86: 664–665.

42 Weiss RB, Gill GG, Hudis CA. An on-site audit of the South African trial of high-dose chemotherapy for metastatic breast cancer and associated publications. J Clin Oncol 2001, 19: 2771–2777.

43 Committee on publication ethics (COPE). Guidelines on good publication practice. Clin Oncol (R Coll Radiol) 2000, 12: 206–212.