Revista Portuguesa de Cirurgia

II Série • N.º 1 • Junho 2007

ISSN 1646-6918

Órgão Oficial da Sociedade Portuguesa de Cirurgia

How to read and analyze the medical literature

Abe Fingerhut, MD, FACS, FRCS (g) [1], *Chadli Dziri MD, FACS* [2], *Ata Toufik MD* [1], *Emanuel Leandros MD, Ph D, FACS* [3]

[1] Department of Surgery, Centre Hospitalier Intercommunal, Poissy, France 78300
[2] Emergency department, Hôpital Charles Nicolle, Tunis, Tunisia,
[3] Department of Surgery, Hippokration University Hospital, Athens, Greece

When caring for their patients, surgeons want to be able to make accurate diagnoses and chose the best treatment option for each particular patient. Such decisions should take into consideration the benefits expected as well as the possible adverse effects associated with the treatment scheme as well as, whenever appropriate, adapting the strategy to the prognostic (risk) factors relative to the patient.

According to Guyatt and Sackett, evidence based medicine (EBM) is about solving such clinical problems, combining the conscientious, explicit, and judicious use or integration of the *best*, currently available external *clinical evidence* (from systematic *research*) with *clinical expertise*, pathophysiological knowledge, and *patient values* in making and carrying out decisions about the care of individual patients".

Intuition, unsystematic clinical experience, and pathophysiological rationale are insufficient tools to use for clinical decision making: EBM leads us to examine the evidence from clinical research to find the best possible solution. With the intention of incorporating the best evidence into their daily practice, surgeons must constantly strive to find the necessary information in the ever-increasing realm of literature that is published today. Sifting the literature for articles that have a high level of evidence is the goal for the physician looking for the answer as to how to make the best therapeutic decision for his or her particular patient. Learning to read and analyze the medical literature with this quest as an objective is the subject of this review: we will concentrate on therapeutic decisions, leaving diagnostic and prognostic considerations (with their specific methodology) for further publications.

When consulting the medical literature to try to answer a clinical question of which therapy is best adapted to the disease of one particular patient, three distinct steps must be addressed: 1) Are the results of the study valid? 2) What exactly are the results? and 3) How can I apply these results to my patients?

Answering the question of whether the ***results of the study are valid*** lays the foundation for credibility: do the results of the study truly correspond to the direction (better or worse) and the magnitude of the underlying true effect found in the study? In other words, do the results found in the study represent an unbiased estimate of the true treatment effect, not influenced by confounding factors that might lead to false conclusions?

What are the results means determining the size and the precision of the treatment effect. Are all the necessary indices included to make the results valid? Last, ***how can I apply these results to patients under my care*** is equivalent to determining, first, whether the results

of the study are generalizable, second, whether they can be "particularized" to the patient you are treating. Last, the reader must also be sure that all the outcome measures have been studied: thus both potential benefits as well as harm, along with the consequences of withholding the patient from treatment, have been analyzed.

The reader must be able to find the type of study easily and without having to dig this information out from the text: usually this should be found in the title or in the key words. The three most often types of studies found in the literature are the ***controlled randomized trial, the cohort study, and the case control study***. ***Controlled randomized trials*** are used when investigators want to assess treatment effects, usually considered to be beneficial. When the investigators want to assess harm, on the other hand, non randomized observational studies can be used, according to whether the patients have been exposed or not to a harmful stimulus, whether as a result of preference, or sometimes as a result of circumstantial chance. Among the non randomized observational studies, when patients are followed forward, and assessed from the time of exposure until the time of the consequences of the exposure (target outcome), this is called a ***cohort study***; when patients are selected once they have the target outcome or not, and researchers look backwards to try to determine the factors of exposure, this is called a ***case control*** study. The qualities and drawbacks of these three types of observational studies are summarized in Table 1. We will endeavor to guide the reader through the analysis of these three types of studies.

How to read and critically appraise a randomized controlled trial (RCT), a cohort, and then a case control study.

In a RCT, the initial step is to determine whether the results are valid. This means studying the patient population to make sure that in the two arms of the study, the "experimental" (the therapeutic arm to be tested) and "control" (the reference or standard) groups, all included patients had similar base-line characteristics, similar prognostic indexes (risk factors), and that the only difference in the two populations concerned whether the treatment (to be tested) was given or not, so that the treatment effect, if found, could be attributed to the treatment and not to some other (confounding) factor. If the number of patients was large enough, most, if not all, the confounding factors should have been more or less evenly distributed between the two treatment arms (control and experimental) and therefore not influence the outcome. The reader should be aware that if this is not the case, the authors could have stratified their randomization to account for the possible effects on the outcome. This pretherapeutic information, describing the patient population, also called demographic data, are most often published in the form of a table, usually the first, listing the characteristics of the two populations under scrutiny.

Were the patients truly randomized? This is occasionally not always as obvious as would seem from simply reading the word "randomized" somewhere in the title or the text. True randomization requires that the allotment sequence (e.g. the choice of administer-

Table 1

DESIGN	STARTING POINT	ASSESSMENT	STRENGTHS	WEAKNESSES
Cohort	Exposure status	Outcome event status	Feasible when randomization of exposure not possible	Susceptible to bias, limited validity
Case-control	Outcome event status	Exposure status	Overcomes temporal delays, may only require small sample size	Susceptible to bias, limited validity
RCT	Exposure status	Adverse event status	Low susceptibility to bias	Feasibility, generalizability

ing one or the other treatments to be compared) be done without any bias. The reader should scrutinize the methods section to make sure that the authors announced that the sequence was: a) consecutive (no eligible patients were left out of the randomization process), b) generated by a method in which it would be impossible to know the next element of the sequence (this means all methods of allotment that rely on date of birth, date of entry to the hospital or order of entry to the study are not valid) and that the actual allotment for a particular patient remains unknown to the patient and to the person administering the care, called "blinding" or "masking" (we will see later on that this is often difficult, if not impossible, in surgical trials). Next, the reader should determine whether the patients were analyzed with the group to which they were allotted, even if for some reason, they did not actually receive the treatment. This is the so-called "intention to treat" principle. The reasons behind the importance of such an analysis are two fold. First, when changing from one arm to other is the result of an unexpected difficulty or pathological finding, analysis of the outcome would then favor the non-difficult or "normal" pathological finding group. Second, unexpected difficulties or pathological findings are part of everyday practice and should be considered to be part of the game.

What was the degree of blinding in the study?

As stated above, the term blinding (or masking) refers to keeping trial participants, investigators (usually health-care providers), and/or assessors (those evaluating and/or collecting outcome data) unaware of the assigned intervention, so that they will not be influenced by that knowledge. Blinding usually reduces differential assessment of outcomes (information bias), but can also improve compliance and retention of trial participants while reducing biased supplemental care or treatment (sometimes called co-intervention). Blinding also ensures that the prognostic factors remained equally distributed in both groups during the conduct of the trial, and the reader should be able to discern whether the patients remained unaware of their allotment all throughout the study, as well as whether the outcome was assessed by some one who was not aware of which arm the patients was allotted to.

Ideally, if the patient, the assessor, as well as the care provider (surgeon) were unaware of the allotted treatment, this would be called a "triple blind" study. Although triple blinding indicates a strong design, trials that are not so should not be rejected automatically or thought to be inferior. Greater credence should be placed in results when at the least, outcome assessment was blinded. In order to assert that blinding was performed correctly, the reader should be able to find in the methods section explicit information as to who was blinded, and how this was done, rather than solely relying on terminology "blinded". If an article claims blinding without any accompanying clarification, however, readers should remain skeptical about its effect on bias reduction. The reader should not naively consider a randomized trial to be of high quality simply because it is "double blind": double blinding is not the *sine qua non* of a randomized controlled trial. Last, one should not confuse blinding with allocation concealment. Such confusion indicates misunderstandings of both.

In surgical trials of surgical technique, however, blinding of the care provider (surgeon) is obviously impossible.

Last, but not the least, the reader should be able to determine if the follow up was complete, and if not, to what degree this (incompleteness) might influence the outcome. A frequently cited example is the follow-up in inguinal hernia studies where the main outcome criterion is recurrence. Patients lost to follow up can of course have died, or moved away, and/or not respond to the recall invitation. But if this were not the case or if this information was not given in the report, it might be that the patient was dissatisfied with the outcome and sought medical advice from another surgeon, or that the patient did not seek medical advice because he or she did not realize that a recurrence has indeed occurred. In order for the results of the study to be considered valid, no more than 10% of patients

should be unaccounted for at the time of assessment. If this is not the case, then the maximal bias or worse case scenario methodologies can be applied. This means considering the outcome of all the patients lost to follow up as a poor result or a failure. If analysis in this manner does not change the results of the study, then it can be assumed that patients lost to follow-up did not to influence (bias) the results. If however, the outcome does change, no valid conclusions can be drawn. Obviously the greater the number of lost to follow up, the less the validity of the study. Such analysis is unfortunately rarely done; the reader should however be aware that in the absence of such analysis, the validity of the conclusions should be mitigated.

What are the results?

This means determining how large the treatment effect was and how precise the results were.

One way of determining how large the treatment effect was is to calculate the absolute risk reduction (ARR) or risk difference. For example if we consider a 10% recurrence rate (x) with one technique of hernia repair (treatment A or control group) compared with a 5% recurrence rate (y) with another technique (treatment B or "treatment" group), the ARR would be x – y = 0.10 – 0.05 = 0.05. We could also express the magnitude of the treatment effect as the RR (the risk of events (recurrence) among patients receiving treatment A relative to that risk (of recurrence) in patients receiving treatment B, or y/x = 0.5/0.10 = 0.5. When expressing results, simple percentages may be misleading (Guyatt JAMA p 97). The most commonly used measure of dichotomous outcomes (e.g. treatment effect yes or no, survival effect dead or alive) is the complement of the RR, called the relative risk reduction (RRR). This is expressed as a percentage (1-y/x) X 100, or in this case, (1- 0.5) X 100 = 50%.

The precise nature (true risk reduction) of the treatment effect is in fact difficult to determine. The best estimate is the observed treatment effect, called the point estimate. As the word "estimate" reminds us, however, we express this (imprecise) fact by calculating *confidence intervals* (CI), that is a range of values within which one can be reasonable confident that the particular population parameter truly lies. Usually the reader finds the 95% CI, a range of values that includes the true risk reduction 95% of the time. The true meaning of CI would be that 95% of such intervals would contain the true value in the population. Conversely, this means that the true RRR will be found outside the CI only 5% of the time, a property which somewhat relates CIs to the conventional level of "statistical significance".

CIs are said to be of quantitative value, as opposed to "p" values, which is a qualitative value, a measure of the strength of evidence against the null hypothesis of "no effect." The p value by itself tells us nothing about the size of a difference or even the direction of that difference. By contrast, CIs indicate the strength of evidence about quantities of direct interest, such as treatment benefit. As such they should be given in the main text and in the abstract of published articles reporting RCTs (Moher Lancet 2001, revised Consort…) and other studies.

CIs are often considered to reflect the clinical significance of results. For instance a CI with a minus value for its lower limit would mean that the treatment effect might even be harmful in some cases, and that the trial under consideration here is of little help to decide whether or not to use the new treatment. The larger the sample size of a trial, the larger the number of outcome events, the narrower the CI and the greater our confidence that the true RRR (or any other measure of efficacy) is close to what has been observed in the study. The adequate sample size then is important: this leads us to consider the number of patients to be included in the study and the risks "alpha", "beta".

The number of patients necessary to include in a RCT can be calculated simply be applying one of several formula which can be obtained in almost any book or computer program. Without going into too many details, let us remind the reader simply that the way the number was calculated should be included in the paper (methods section). The calculation should take into account four parameters: the alpha and beta

risks, the delta (difference expected) and the standard deviation (if means). The four parameters are the alpha and beta risks, plus the percentages of the experimental and control groups for qualitative variables.

When we look at the clinical significance again, in a positive study, if the *lower* limit of the CI is still consistent with the RRR considered sufficiently effective to recommend this treatment to your patient, then the number of patients enrolled was adequate. If on the other hand, this lower limit no longer clinically relevant, then the trial result can not be recommended even though the results might be statistically significant. In a negative study, on the other hand, a look at the *upper* limit determines whether this limit is clinically relevant. If so, the reader can say that not only has the study failed to show that the experimental treatment is better than the control modality, but also that the trial failed to prove that it is not. Absence of evidence is not proof (evidence) of absence (of treatment effect).

The usual means of comparing the differences (statistical tests to be used), between the two groups, whether comparing the demographic data or the results, depends on whether the results are continuous (numerical values on a scale) or categorical (yes/no, dead/alive,…). Another consideration to analyze is if the data given are "normally" distributed or not. Most statistical tests are either parametric (relying on a specified data distribution) or nonparametric (not relying on a specified data distribution). Many parametric tests rely on the "normal distribution", that is a distribution that can be represented by the symmetric, bell-shaped Gaussian distribution curve. On the other hand, a nonparametric test is generally preferred if the distribution of data is clearly non-Gaussian. One commonly observed error is to see durations (operation, length of stay, period of recuperation…) expressed as means with standard deviations. While age, when the number of patients is less than 30, may or may not be distributed according the Gaussian, or "normal" curve, durations are never "normal". The reader should realize that whenever it is unclear whether or not the data are normally distributed (and therefore allow analysis by a parametric test), it is usually better to see a non-parametric test being used because the latter yields slightly more conservative "p" values.

When looking at normally distributed "continuous" parameters, the most often used statistical test is the Student t test. If the distribution is not normal, however, the Mann-Whitney-U Test should be used.

As an example, if one wants to compare the corresponding values of blood sugar between two independent (unpaired) samples of patients with and without diabetes, the t test is appropriate if the distribution is normal (mean values can be used), and the Mann-Whitney test should be used when the distribution is not normal (median values (*not* means) should be used for the comparison).

As another example, if one wants to measure the diameter of a rectal tumor in patients with rectal cancer before and after neoadjuvant radiotherapy, for instance, the t test can be used when the sample is unique (paired data), when there are at least 20 measures, and when the distribution is normal. Otherwise, the Wilcoxon matched paired test should be used (once again medians are more appropriate than means when the data is not Gaussian).

For "categorical" outcomes, the most often used statistical test is Pearson's Chi squared test, or when the number of measurements are less than 20, Fischer's exact test should be used. When data are paired, McNemar's test should be used.

To compare the average number of sampled lymph nodes in three groups of patients undergoing different resection methods of pancreatoduodenectomy for pancreatic cancer (whether paired (one sample) or unpaired (two different samples)), the one way analysis of variance (ANOVA) should be used for parametric data, and the Kruskal-Wallis test for non parametric data.

Last, the reader should be able to discern whether the outcome of the study can apply to the patients he or she sees and has to treat. Often the patient in question differs more or less from the patient(s) enrolled in the trial you want to use for your decision: for instance, your patient might be older, sicker, or may have co-morbidity or another condition which was

not included (excluded) in the trial under scrutiny.

Even if your particular patient would have qualified for entry into the study, the reader should remember that treatments are not uniformly effective in each and every individual. The "overall" treatment effect is in fact an "average treatment effect" and as such, applying the results of the trial might expose a particular patient to harm or extra costs without any benefit.

One often-criticized characteristic of clinical studies is when they are multicenter studies. Multicenter studies are performed mainly because large numbers of patients can be included in a short period of time. Moreover, their multicenter attribute, allows the results to be generalized. However, these same multicenter studies are often rebutted because the lever of expertise or, and often, by consequent, the level of care, is not equal, leading to results that are often less favorable than in monocenter studies of the same treatment. Arguments in this direction are usually addressed in the discussion section of the article.

When the characteristics of the patient considered for therapy are not exactly the same as those (inclusion) criteria used for the trial, the treatment may still be applied, if there is no compelling reason not to do so.

In observational studies, the reader has to ask the same questions: " are the results valid", "what are the (exact) results", and "how can I apply them to my patient(s)"?

To ascertain whether the results are valid, the same reasoning leads to reader to assess whether the patients similar from the start, and were their basic demographics and prognostic factors similar.

Cohort studies: as stated above, consist of identifying exposed and nonexposed patients, following them forward in time, and then monitoring their outcome. They are particularly useful when assessing rare events, such as possible harm, because, a RCT under these conditions usually requires many subjects, and when they study harm, ethically questionable or even unrealizable (subjects would have to be informed of the possible harm). The danger here is that the two groups (exposed and non exposed) may begin the study with different risks for the outcome criterion. Furthermore, there are usually other associated reasons (potential confounding factors) that may influence the decision to prescribe one treatment or the other. In this case, these differences must be recorded and analyzed, and then the reader has to be able to determine whether either the two groups were similar as concerns all the factors excepting the exposure or else, find that the authors used appropriate statistical techniques to adjust for these differences. The most widely used test for this is the kappa correlation test, a measure of chance-corrected agreement. The closer the kappa value is to 1.0, the better the agreement. Kappa values over 0.75 are generally accepted to be of high degree of agreement. However, to the contrary of randomized trials, where unknown confounding factors should be evenly distributed by chance after randomization, in cohort studies, this type of bias may still exist, but go unrecognized. Here the strength of inference deriving from a cohort study will always be less than that of a RCT.

Case-control studies

When the event is rare or especially when they take a long time to appear, an alternative technique of investigation is the case-control study. In this type of investigation, patients who have already developed the target outcome are compared to a group of persons, who, as a group, are similar in demographics such as age, sex, and prognostic factors, but who have not developed the target outcome. The reader should be able to find an evaluation of the relative frequency of exposure to the putative agent present in either group, and discern whether the authors adjusted for differences in the known and measured prognostic factors. In this respect, all possible confounding factors should have been counted and analyzed.

In both these types of observational studies, the reader now has to be able to discern whether the exposed patients were equally likely to be identified in the two groups. In order to avoid inherent biases which would appear if the subjects were asked about the possibility of exposure (recall bias) or because of

the insistence or motivation of the person who interrogated the patient (interviewer bias), blinding of the participants and interviewers should be employed. Another bias inherent to gathering of information when the interviewer is not blinded is his or her greater perspicacity of detecting risk factors and associated disease in the experimental or exposed group, and therefore creating what is called a surveillance bias.

As with RCT, the reader must assure himself or herself that follow-up was as complete as possible. Otherwise, of course, in the same manner, the missing patients should be accounted for in one way or another.

To the question of "what are the results", the reader has to now assess the strength of association between the exposure and outcome and then determine how precise the estimate of the risk.

In ***cohort studies***, the relative risk is calculated as before, expressing the increased (ratio greater than 1) or decreased (ratio less than 1) risk of developing the outcome measure when the subject is exposed compared to when he or she is not exposed. In ***case control studies***, the relative risk cannot be calculated because the number of cases and controls, the proportion of persons with the outcome, has been chosen by the investigator. In these types of studies, the reader should find ***odds-ratios***, the odds of a case patient being exposed, divided by the odds of a control patient being exposed. As stated above, inference of a relationship between exposure and the outcome is weaker in observational studies than in RCT. However, two characteristics add weight to the inference, a high RR or OR, on one hand, and second, when several measurements have been made, a strong dose response relationship (increase in the outcome measure proportional to the increase in exposure), on the other.

The precision of the estimate of the risk is given by the width of the confidence intervals around that estimate. In a study where an association has been shown, the lower limit of the CI of the estimate of the RR associated with the risk of the exposure determines the minimal estimate of the strength of the association. On the other hand, in a negative study, the upper limit of the CI tells the reader how big the risk may be in spite of the negative result of the association (no statistical signification found). The reader should be aware that many papers confuse RR and OR. RRs are easier to conceptualize and therefore are likely to be used instead of OR, more difficult to apprehend. The error, however, is acceptable, provided that the event rate is low.

Last, the reader must now attempt to determine first whether the results of the observational study can be generalized to the overall population, and then whether the results are applicable to a patient he or she might encounter. As before, the reader needs to be able to find sufficient details in the paper as to the patient population, in order to determine if his or her patient fits the patient profile found in the study.

In conclusion, a complete and step-by-step assessment ("critical appraisal") of all scientific papers is necessary to ascertain the validity, the credibility and the generalizibility of the information. This is a prerequisite before the reader can draw any conclusions or infer any associative properties. Any empirical observation

Meta analysis of randomized controlled trials (RCT)
Systematic review of RCT
Single RCT
Systematic review of observational studies addressing patient-important outcomes
Single observational study addressing patient-important outcomes
Physiological studies
Unsystematic clinical observations

Table 2 – **A hierarchy of strength of evidence for treatment decisions**

about the apparent relation between events constitutes potential evidence. The reader must then be able to hierachize this evidence according to what are called "levels of evidence". The Canadian Task Force on the Periodic Health Examination first brought to the attention of the medical community five grades of evidence, based on study designs. The hierarchy of evidence for treatment effects can be found in table 2. This hierarchy, however, is not absolute. If treatment effects are large enough and consistent, for example, observational studies may provide compelling evidence, even in spite of existing RCT.

Once again, it is not just because the study is announced as a randomized controlled study, not just because the results are esthetically or astutely tabled, not just because the authors state, that "according to the results of their study, they can conclude that…." and last, not just because the numbers are associated with "statistically significant" "p" values or confidence intervals that the reader can take the results of the paper as proven. It is the reader's task to analyze the data according to the outlines given here (critical appraisal). Then and only then can the reader be confident that the results announced can be applied to a patient under his or her care with benefit and without creating any adverse risk.

SELECTED READING

Altman D. Why do we need confidence intervals? World J Surg 2005; 29: 554-6

Altman DG, Bland JM. "Absence of evidence is not evidence of absence" BMJ 1995; 311: 485

Canadian Task Force on the Periodic Health Examination. The periodic health examination. CMAJ. 1979; 121: 193-54

Dziri C, Fingerhut A. What should surgeons know about evidence-based surgery. World J Surg 2005 ; 29 : 545-6

Greenhalgh T. How to read a paper: statistics for the non-statistician. I Different types of data need different statistical tests. BMJ 1997 ; 315: 364-6

Greenhalgh T. How to read a paper: Statistics for the non-statistician II "Significant relations and their pitfalls". BMJ 1997; 315: 422-5

Greenhalgh T. How to read a paper: getting your bearings (deciding what the paper is about). BMJ 1997; 315: 243-6

Greenhalgh T. How to read a paper: assessing the methodological quality of published papers. BMJ 1997; 315: 305-8

Guller U, DeLong ER. Interpreting Statistics in Medical Literature: A Vade Mecum for Surgeons JACS 2004 ; 198 : 441-53

Guyatt G, Rennie D. Users' Guided to the Medical Literature. Essentials of Evidence-based Clinical Practice. JAMA Chicago 2002

Haynes ACP Journal Club 1996; 124: 14-5

Hiatt RA , Krieger N, Sagebiel RW, Clark WH Jr, Mihm MC Jr. Surveillance bias and the excess risk of malignant melanoma among employees of the Lawrence Livermore National Laboratory. Epidemiology 1993 ; 4 : 43-7

Millat B, Borie F, Fingerhut A. Patient's preference and randomization: new paradigm of evidence-based clinical research. World J Surg 2005 ; 29 : 596-600

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001: 357: 1191-4

Robbins AS, Chao SY, Fonseca VP. What's the relative risk? A method to directly estimate risk ratios in cohort studies of common outcomes. Ann Epidemiol 2002 ; 12 : 452-4

Shulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002; 359: 696-700

Address for correspondence and reprints:

A Fingerhut
Centre Hospitalier Intercommunal, 78300 Poissy France
abefinger@aol.com