Acta Orthop. Belg., 2004, **70**, 91-97

REVIEW ARTICLE

Evidence-based medicine in orthopaedics

Olivier BARBIER, Michel HOOGMARTENS

Evidence-based medicine (EBM) is medicine based on the sum of personal clinical experience and clinical studies with the best possible design (preferably, but not necessarily, randomised controlled trials or RCTs), while focusing on the expectancies of patients and institutions.

INTRODUCTION

Many of our medical interventions are still based on feelings or traditional therapies. For example, despite the scientific proof that preoperative shaving, the use of skin drapes or wound drainage do not lower but rather increase infection rates, those practices remain in use in many hospitals. But something may change with the development of “evidence-based medicine”. This term was first used in the early nineties by Guyatt (*16*), and became widespread through a series of articles published by the Evidence-Based Medicine Working Group, from 1992 on (*14*).

As a *first innovation*, this new philosophy added to the clinician’s experience the importance of strictly applied study protocols, like the randomised controlled trial (RCT) (*17*). Of course, RCTs were known since years. For instance, the first well-documented randomised controlled trial was reported in 1948 : it compared streptomycin with the classical treatment for pulmonary tuberculosis. Randomisation, where hazard decides about grouping of patients, is the only method that allows to maximally exclude the difference between two comparison groups, as to known and unknown prognostic factors. Randomisation automatically includes the use of a control group.

A second innovation was the fact that the new approach finally showed respect for the patient’s needs and values, and later that of certain institutions, such as the health systems, concerned about the cost/benefit ratio. This led to the definition of new outcomes of interest in the clinical studies. So one can rightfully state that something changed.

Advocates of “evidence-based medicine” classify studies according to “grades of evidence” on the basis of their scientific level (*12*).

The lowest level of evidence (level 5) is the *expert opinion* (*10*).

The *case report* (about one single patient) and the *case series* (also retrospective, but concerning more patients) reach a slightly higher level (level 4) ; they fill most pages in our medical journals. For example, one looks at all the hip prostheses inserted between 1985 and 1995. A control group is normally not involved, but sometimes a historical control group can be used ; this is,

From Saint-Luc University Hospital, Brussels and Pellenberg University Hospital, Leuven, Belgium.

Olivier Barbier, Associate Surgeon-in-Chief.

Department of Orthopaedics, Saint-Luc University Hospital, Université Catholique de Louvain, Brussels, Belgium.

Michel Hoogmartens, Associate Professor emeritus.

Department of Orthopaedics, University Hospital, Katholieke Universiteit Leuven, Pellenberg, Belgium., Pellenberg, Belgium.

Correspondence : Olivier Barbier, University Hospital Saint-Luc, Avenue Hippocrate 10, B-1200 Bruxelles, Belgium. E-mail : olivier.barbier@orto.ucl.ac.be.

however, a doubtful procedure. These studies are important to generate clinical questions and hypotheses but are inappropriate to conclude about the efficacy of a treatment.

The *case-control study* comes next (level 3) : it is also retrospective, but compares two non-randomised groups : one with and one without the outcome event. So the outcome event is already known, and the study looks for the influence of treatment or risk factors. For instance a group of well-healed fractures is compared with a group of nonunions (both known outcomes) to find out if the second group included more smokers.

A *cohort or observational study* reaches a higher level : level 2. In contrast with the case-control study the treatments or the risk factors are known, while the outcomes are not. For instance, one follows thousands of menopausal women with and without hormone replacement, looking for hip fractures to occur. So this prospective or retrospective cohort study follows groups of well-defined people, like Roman cohorts, exposed or not exposed to the treatment, for the development of the outcome of interest. There is no random assignment to treatment arms. A *prospective cohort study* is one where the interventions are identified prior to study onset, and the outcomes of interest occur after a specified follow-up period. A *retrospective cohort study* is one where the outcomes have already occurred by the time the study is initiated.

The *randomised controlled trial* (RCT) reaches a 1B level, especially when the patients are blinded (single blind study) and even more so when both the patients and the investigators are blinded (double blind study).

Level 1 is reached by the *systematic review*, which combines several randomised studies on a specific issue, and by the *meta-analysis*, which mathematically combines several randomised studies on a particular issue.

When most aspects of a given problem, such as hip fracture, have been studied at the highest levels, it is possible to write *guidelines* : operation within 24 hours, internal fixation in younger patients, arthroplasty in the very old, anti-thrombotic prophylaxis.

FIRST INNOVATION : A DISTINCT PREFERENCE FOR THE RCT

The first question the critical reader should ask about an RCT is whether the characteristics of the study are clearly defined regarding the inclusion and exclusion criteria, the modalities of treatment and the outcomes studied. The inclusion and exclusion criteria allow the clinician to have an idea about the general applicability of the outcome to his own patients. Furthermore, a perfect definition of the treatment limits possible disagreement due for example to a different definition (*1*) for two teams, of open reduction and internal fixation as treatment for intra-articular calcaneal fractures (*11*). Especially, the surgeon must be aware of a learning curve or an exceptional technical skill of the team performing the study, before he can generalise the results to his own daily practice. And finally, the outcomes studied have to be clearly defined in order to permit statistical analysis.

The next question concerns randomisation, in order to avoid a selection bias (*13*). Alternate assignment to groups, or assignment based on chart number etc…still allow for bias, as the randomisation is not concealed : the investigator knows beforehand to which group the patient will belong, and he can manipulate the allocation. In fact the investigator should first decide if the patient is suitable for the study, and only then open an opaque numbered envelope that contains the specification of the group to which the patient will belong. Even better is the telephonic system, where a computer notes the name of the patient and tells the investigator what to do. Another good system is medication that looks alike for the treatment group and for the control group. However, many situations limit the random assignment to a group. Trials assessing surgical interventions can compare the intervention with placebo (ethically often not feasible) (*20*), with conservative treatment (no surgery), or with another surgical intervention. Moreover, in orthopaedics the patients will often be reluctant to accept that hazard will decide if they will receive conservative or surgical treatment. Patients in the various groups must be similar as to “known” prognostic factors. It makes no sense to compare two methods of wrist arthrodesis if one group contains most of the rheumatoid patients and the other one those with osteoarthritis. The rheumatoid and osteoarthritic cases should be studied separately in the two groups : stratification, before randomisation. Of course, very big groups will largely avoid this kind of mistakes, but stratification would still be useful. Big groups will also tend to correct for "unknown" prognostic factors. The surgeons who perform the operations must be proficient in the techniques which are compared. But if two surgeons A and B are each familiar with only one of two techniques, respectively A' and B', it is possible to randomise the surgeons. Ideally, this should be done in the same hospital so that the aftercare is the same. Blinding is another important feature. It is difficult to blind the surgeon, unless the operation consists of, for instance, the injection of chymopapain versus placebo. But patients, data collectors and outcome assessors should be blinded as much as possible. Otherwise they can transmit their enthusiasm for the new treatment to the patient.

The third point to analyze is the follow-up, which should be as complete as possible. Even patients who withdraw from the study must be analysed as members of their original group. This according to the "intention-to-treat" principle. Less than 5% drop-outs is ideal, and 6 to 20% is the grey zone. When there are many drop-outs, a "worst case" analysis should be done : all the drop-outs are considered as having the unexpected issue. This means that every patient lost from the group with the best results is considered to have a bad outcome, while every patient lost from the group with the worst results is considered to have a good result. Completeness also means that the follow-up is long enough to detect the outcome. Treatments for lumbar spondylosis and for herniated disk need at least two years of follow-up.

Ideally, the groups should include all the cases in the world that would be eligible. In that ideal situation, simple descriptive statistics summarising data (mean, median, variance,…), and presentation in a digestible form such as graphics would be sufficient. Unfortunately, most often the groups are just samples, which means that probability comes in. Indeed, as samples only partially reflect the reality, one can merely state that it is probable to a given degree that the findings are correct, helped in this task by the second branch of general statistics, the inferential statistics, which are based on hypothesis testing. Therefore, a few words are necessary about statistics and the most common mistakes that occur in this matter.

Type I or alpha error. It is common knowledge that most investigators try to find a difference between two groups, and that they hope to reject the null hypothesis (*9*), which states that there is no difference at all, unless a sufficient weight of evidence indicates otherwise. By convention, most authors accept a 5% risk (alpha = 0.05) that the results of a study are thought to be true, while in fact they are attributable to chance. This is called a *Type I or alpha error.* If the probability p is lower than 0.05, then we reject the null hypothesis. The value of p is a somewhat arbitrary decision, but 0.05 is mostly used and generally works well in practice. So the value of p that determines whether we accept or reject the null hypothesis is commonly called alpha. If one accepts a 5% risk (alpha = 0.05), then one out of 20 studies on the same subject will be positive by mere chance. But there is more. If one makes enough comparisons, one can almost be certain to find one or more that exhibit a significant difference. For instance, if one looks for the correlation between low back pain and multiple factors like obesity, smoking, job satisfaction, spondylolisthesis and others, there is a great risk that correlations will be found by mere chance, if one accepts an alpha level of 0.05, which is valid only when only a single factor is tested. To compensate for this, we can use the so-called Bonferroni adjustment, or other adjustments : the alpha threshold will need to become 0.001 for instance, instead of 0.05. Because of this penalty, it is important to study as few factors as possible.

Type II or beta error. Many studies conclude that alpha was > 0.05 and, consequently, that the difference between groups was not significant. However, this is not necessarily true (*6*). It may well be that the samples were too small, and that a difference would have been found if the samples had been larger (*19*). This is called a *Type II or beta error.* Beta is the risk of accepting a Type II error, just like alpha is the risk of accepting a Type I error. Beta is mostly set, by convention, at 0.20, which means that one is ready to accept a 20% chance that the study concludes that there is no difference between groups, while in fact there is a real difference. The statistical power is defined as 1 – beta, which is equal to the chance not to commit a Type II error, or equal to the chance to demonstrate a difference between two treatments when one actually exists. Power depends in the first place on the magnitude of the samples, but also on the variability of the results and on the difference between groups. The variability of the results, often expressed as the standard deviation, plays a role, because a pronounced variability means that the individual results are quite remote from the mean, so that more measurements are necessary to obtain an idea about the true difference. A large difference between groups will, of course, require smaller samples : it is easier to see a difference when it is striking. It is good to note that the desired difference between groups must be fixed *before* the study, and that it is based on clinical feeling or preliminary studies, frequently with a lower scientific level. In other words, the investigator must state beforehand that he considers a difference between the surgical treatment A and the conservative treatment B as worthwhile, if treatment A improves the manual ability in rheumatoid hands with more than 1 unit on a functional scale (ABILHAND, for instance) (*21*) in the activities of daily living. By performing power and sample-size calculations prior to conducting a trial, the investigators can reduce type-II error rates.

Most studies completely neglect the type II error (*19*). One of the reasons is the fact that enormous samples are necessary to avoid it. A type I error is made less frequently. Another means to avoid a *type I* or a *type II* error is to observe the confidence interval of the results. To calculate this interval often remains the domain of the statistician, although a computer program can help in this task (*2*). The convention of using the value of 95% is arbitrary, meaning that we can be 95% sure that the true value lies within the limits of the confidence interval which is closely related to the conventional level of statistical significance $p < 0.05$. In contrast to the p value which is only a measure of the strength of evidence against the null hypothesis, the confidence interval indicates the size and the direction of the difference. To avoid a *type I* error, in a positive study, the clinician can look at the lower number or boundary of the confidence interval and decide if this value is still of clinical significance. To avoid a *type II* error, in a negative study, one needs to examine the upper number or boundary of the confidence interval. If this value would, if true, be clinically important, then the study has failed to exclude an important treatment effect (*7*).

After all these precautions we can suppose that the results of a given study are valid.

SECOND INNOVATION : PATIENTS AND INSTITUTIONS BASED OUTCOMES

In the past, most studies were mainly interested in outcomes such as joint mobility, walking distance, radiological axes, and others, with little respect for the opinion of the patient. These outcomes are now considered as first level outcomes, simply based on *function*, according to the International Classification of Functioning, Disability, and Health, known as ICF (World Health Organisation) (*28*).

Today, second level outcomes focus on the *disability* experienced by the patient himself. Questionnaires allow him to tell, without the intervention of the clinician, if he is still able to perform the activities of daily living, how he feels psychologically, or how much pain he experiences.

The third level outcomes now give an idea about *health*-related quality of life (HRQL). The SF (Short Form)-36, consisting of 36 questions, is one of the most widely used instruments (*27*). It summarises the patient's perception of his physical and mental health.

The questionnaire is considered an instrument and should possess the scientific characteristics of reliability, validity, responsiveness, and appropriateness (*25*). The patient evaluates his own condition by answering a questionnaire, without the intervention of the clinician : one of the most striking tools included in the evidence-based medicine.

The patient based outcomes can be classified into four types (*4*) :

A. Generic Measures

The SF-36 is the most commonly used generic measure.

B. Specific Measures

1. *Region-specific measures* cover a whole region, for instance the upper limb, like the Disabilities of the Arm, Shoulder and Hand outcome measure (*3*).
2. *Joint-specific* or *Disease-specific measures* evaluate a single joint or disease. Examples are : the WOMAC (Western Ontario McMaster Osteoarthritis Index) (*5*) for hip and knee ; the Roland-Morris Questionnaire (*22*) and the Oswestry Disability Index for low back pain (*15*).
3. *Patient-specific Measures*. Strangely enough the patient chooses himself, on a customised questionnaire, the questions that are applicable to his condition : for instance either the questions on light activities or those on heavy activities (*29*).

C. Utility Measures

Utility measures often assess a person's preference for a state over an other.

The EQ-5D or Euro-QOL uses a scale from 0 to 1 (where 0 means death and 1 excellent health), and sometimes even negative marks below zero to describe a situation worse than death (*26*). The final score is used in the construction of quality-adjusted-life-years (QUALYs) where a year in a higher-quality health state contributes more to the outcome than a year in a poor-quality health state. The QUALYs are useful for health-economists who calculate the cost per quality of life year gained by a given therapy. Institutions such as health systems and managed care systems are interested in these matters. This had led to the objection that evidence-based medicine merely serves the purpose of hospital managers.

D. Other outcome measures

They estimate pain (frequently by a visual analog scale), satisfaction (which is a multifactorial component), or work-related disability (with the Work Limitations Questionnaire) (*18*).

In practice most researchers use a combination of a Generic and a Specific outcome.

Now that we have given the patients, who took part in the randomised study, an opportunity to express their personal feelings about the result, we are still faced with the problem of applying that result to the patient we have to treat. Indeed, the individual patient with ischaemia of the lower limb will be glad to hear that a randomised study (*24*) about the infection rate after amputation pleaded significantly for the use of preoperative antibiotics, but he will not be interested in p being smaller than 0.005. He wants to know "how far" the risk of an infection is reduced by these antibiotics, in other words he wants to have an idea about the magnitude of the treatment effect (*23*). He leaves the validity of the study to the clinical researcher, but he wants to know how important the findings are. The clinician can tell him that the risk for sepsis was reduced from an absolute risk of 39% to an absolute risk of 17%. Furthermore, that the absolute risk reduction is the difference between the absolute risks, or 39%-17% = 22%. Now the clinician can inverse that absolute risk reduction of 22%, so that it becomes 1/22% or 1/0,22 or 4,5 say 5. This means that he needs to treat 5 more ischaemia patients with preoperative antibiotics in order to prevent one more infection, *if he adopts the same timing of the therapy as in the trial.* In short, the *number needed to treat*, or NNT, is a term that becomes more and more classic in evidence based medicine (*23*), revealing clinical significance of statistics.

Again, the statistician will be charged with the task to calculate the 95% confidence interval of the NNT, or the limits within which the true NNT lies 95% of the time. This will tell us how precise the NNT is as an estimate of the treatment effect.

Finally, the clinician must check if his patient is too different from the patients in the study. The above mentioned study excluded patients with a temperature above 38° Celsius, which means that a patient with a temperature of 40° is likely to have a worse outcome, even with preoperative antibiotics. The clinician will also study the feasibility of the treatment (is the patient allergic to antibiotics ?), the possible harms for the patient (shock in case of allergy) and the expectations of the patient.

HOW TO FIND THE EVIDENCE BASED LITERATURE ?

One can distinguish three levels in the obtainment of evidence-based information :

– "randomised controlled trials" focusing on a very specific question, for instance : "Are infiltrations with steroids helpful for the treatment of tennis elbow ?". Of course, the search engine prefers to receive the question formulated as "steroids *and* tennis elbow" : "*and*" is a so-called Boolean operator and makes sure that only steroid therapy for the treatment of tennis elbow will be considered, thus avoiding several thousands of results for "steroids", and several hundreds for "tennis elbow".
– "systematic reviews", compiling the data of several randomised controlled trials, sometimes as a mathematical meta-analysis if all the conditions are fulfilled.
– "guidelines", reflecting the whole existing philosophy about a more general problem, such as "lumbar fusion", based on randomised controlled trials.

The Journal of Bone and Joint Surgery, American volume, started with the publication of evidence-based articles in June 2000. Since 2001 they appear on a regular basis in the February, May, August and November issues.

However, the electronic media, such as the computer disk, the CD-ROM and internet permit a much wider access to the evidence-based literature. The Ortholine CD-ROM can be purchased through the Journal of Bone and Joint Surgery : http://www.jbjs.org.uk/subs__cd.htm. Pubmed (http://www.pubmedcentral) is free of charge, and very practical for the orthopaedic surgeon who wants to obtain much information in a few minutes. If he prefers to limit his search to systematic reviews, it is sufficient to click on clinical queries » systematic reviews. A search term like "glucosamine and osteoarthritis" will yield several systematic reviews. The Cochrane Library, called after the British epidemiologist Archie Cochrane (1909-1988), a forerunner, contains more than 200.000 randomised clinical trials. Systematic reviews of randomised controlled trials across all areas of health care are prepared by international collaborative review groups. Access is free of charge in England and Wales, Ireland, Finland, Norway and Australia ; in Belgium a fee of 50 euro is required, except in certain medical libraries : www.cebam.be » choose your language» virtuele bibliotheek/ biblioth. virtuelle » Cochrane database of systematic reviews. Somehow, free access to some abstracts of the Cochrane reviews is possible via www.update-software.com » The Cochrane library » Abstracts of Cochrane reviews. Guidelines can also be obtained, free of charge, from http://www.guidelines.gov : search terms like "hip fracture" or "open fracture" produce a survey of the state of the art. A researcher, who wants to start a clinical trial, can find out if similar trials are going on : http://ClinicalTrials.gov.

THE PRACTICE OF EVIDENCE BASED MEDICINE

One can distinguish five steps to practice EBM (23). A question about a patient serves as the starting point. The best answer is searched through studies of the highest possible level, in the second step. The results of this search are critically appraised for their validity and importance in the third step, to be finally applied to the particular patient in the fourth step. The reflection to improve this process constitutes the fifth step. It is essentially a way to rationalise but not to limit the medical practice.

Moreover, every clinician, who takes part into studies of the highest level, helps to promote Evidence Based Medicine.

CONCLUSION

Evidence-based medicine is a fresh way of rational thinking in the approach to the truth in clinical problems, with the patient's needs and values as end-points. But the truth may hurt. It is not agreeable to hear that microdiscectomy is not significantly better than standard discectomy, and that instrumented lumbar fusion is not significantly better than simple fusion. Self-delusion should be banned from medicine.

REFERENCES

1. **Alonso J, Buckley R, Benirschke SK.** Back page debate : calcaneal fracture. *AO Dialogue* 2003 ; 16 : 37-40.
2. **Altman DG, Machin D, Bryant TN, Gardner MJ (eds).** *Statistics with Confidence.* 2nd ed. British Medical Journal, London, 2000, (including software).
3. **Beaton DE, Katz JN, Fossel AH *et al.*** Measuring the whole or the parts : validity, reliability and responsiveness of the DASH Outcome Measure in different regions of the upper extremity. *J Hand Ther* 2001 ; 14 : 128-146.
4. **Beaton DE, Schemitsch E.** Measures of health-related quality of life and physical function. *Clin Orthop* 2003 ; 413 : 90-105.
5. **Bellamy N, Buchanan WW, Goldsmith CH *et al.*** Validation study of WOMAC : a health status instrument for measuring clinically-important patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. *J Orthop Rheum* 1988 ; 1 : 95-108.
6. **Bernstein J, McGuire K, Freedman KB.** Part II. Statistical issues in the design of orthopaedic studies. *Clin Orthop* 2003 ; 413 : 55-62.
7. **Bhandari M, Guyatt GH, Swiontkowski MF.** User's guide to the orthopaedic literature : how to use an article about a surgical therapy. *J Bone Joint Surg* 2001 ; 83-A : 916-926.
8. **Bhandari M, Richards RR, Sprague S, Schemitsch EH.** The quality of reporting of randomized trials in the Journal of Bone and Joint Surgery from 1988 through 2000. *J Bone Joint Surg* 2002 ; 84-A : 388-396.
9. **Bhandari M, Whang W, Kuo JC, Devereaux PJ, Sprague S, Tornetta III P.** The risk of false-positive results in orthopaedic surgical trials. *Clin Orthop* 2003 ; 413 : 63-69.
10. **Brighton B, Bhandari M, Tornetta P, Felson DT.** Part I. Methodological issues in the design of orthopaedic studies. *Clin Orthop* 2003 ; 413 : 19-24.
11. **Buckley R, Tough S, McCormack R *et al*.** Operative compared with nonoperative treatment of displaced intra-articular calcaneal fractures : a prospective, randomized, controlled multicenter trial. *J Bone Joint Surg* 2002 ; 84-A : 1733-1744.
12. **Concato J, Shah N, Horwitz RI.** Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000 ; 342 : 1887-1894.
13. **Devereaux PJ, McKee MD, Yusuf S.** Methodological issues in randomized controlled trials of surgical interventions. *Clin Orthop* 2003 ; 413 : 25-32.
14. **Evidence-Based MedicineWorking Group.** Evidence-based medicine : a new approach to teaching the practice of medicine. *JAMA* 1992 ; 268 : 2420-2425.
15. **Fairbank JCT, Couper J, Davies JB *et al.*** The Oswestry low back pain questionnaire. *Physiotherapy* 1980 ; 66 : 271-273.
16. **Guyatt GH.** Evidence-based medicine. *ACP J Club* 1991 ; 114 : A16.
17. **Jadad AR, Rennie D.** The randomized controlled trial gets a middle-aged checkup. *JAMA* 1998 ; 279 : 319-320.
18. **Lerner D, Amick III BC, Rogers WH *et al.*** The work limitations questionnaire. *Med Care* 2001 ; 39 : 72-85.
19. **Lochner HV, Bhandari M, Tornetta III P.** Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg* 2001 ; 83-A : 1650-1655.
20. **Moseley JB, O'Malley K, Petersen NJ *et al*.** A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med* 2002 ; 347 : 81-88.
21. **Penta M, Thonnard JL, Tesio L.** ABILHAND : a Rasch-built measure of manual ability. *Arch Phys Med Rehabil* 1998 ; 79 : 1038-1042.
22. **Roland M, Morris R.** A study of the natural history of back pain : Part I : development of a reliable and sensitive measure of disability in low-back pain. *Spine* 1983 ; 8 : 141-144.
23. **Sackett DL, Straus SE, Richardson WS *et al*.** *Evidence-based Medicine. How to Practice and Teach EBM.* Churchill-Livingstone, Edinburgh, 2nd edition, 2000, p 111.
24. **Sonne-Holm S, Boeckstyns M, Menck H *et al.*** Prophylactic antibiotics in amputation of the lower extremity for ischemia. *J Bone Joint Surg* 1985 ; 67-A : 800-803.
25. **Szabo RM.** Outcomes assessment in hand surgery : when are they meaningful ? *J Hand Surg* 2001 ; 26-A : 993-1002.
26. **Tosteson AN.** Preference-based health outcome measures in low back pain. *Spine* 2000 ; 25 :3161-3166.
27. **Ware JE, Snow KK, Kosinski M, Gandek B.** *SF-36 Health Survey : Manual and* I*nterpretation Guide*. Boston : Health Institute, New England Medical Center, 1993.
28. **World Health Organization.** *International classification of functioning, disability and health (ICF).* WHO, Geneva, 2001.
29. **Wright JG, Young NL.** The patient-specific index : asking patients what they want. *J Bone Joint Surg* 1997 ; 79-