“Evidence-based medicine” is a popular concept that simply means basing clinical decisions on best available scientific evidence. The idea appears very sensible and attractive – who would not want to be treated according to the best scientific evidence? As usual, complexity lurks just below the surface.

The natural next question is – what is good scientific evidence? Medical literature is huge, full of controversies, and highly variable quality of data and presentation. In order to make sense of the mess, various organizations around the world tried to classify evidence by presumed quality. In the United States, in a text published in 1989, the U.S. Preventive Services Task Force devised one such system of ranking medical literature:

- Level I: Evidence obtained from at least one properly designed randomized controlled trial.
- Level II-1: Evidence obtained from well-designed controlled trials without randomization.
- Level II-2: Evidence obtained from well-designed cohort or case – control analytic studies, preferably from more than one center or research group.
- Level II-3: Evidence obtained from multiple time series with or without intervention. Dramatic results in uncontrolled trials might also be regarded as this type of evidence.
- Level III: Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

Presumably, a randomized controlled trial (RCT) avoids key sources of bias associated with other study designs, and thus provides more reliable data – justifying higher level of trust. The idea behind a randomized blinded trial is solid – the design helps eliminate biases in treatment administration and data analysis, conscious or not. The paradigm proved its potency in experiments on laboratory animals, and in some large – scale clinical trials of medical treatment. Despite occasional difficulty with variable populations and inclusion criteria, medical treatment studies allowed real blinding and real control. However, the classification system that permanently elevates an RCT above other types of evidence is empirical; it is the result of an expert panel opinion – level III evidence according to the classification itself. This system of ranking articles was used by other medical organizations and medical journals to generate “classes” of evidence, where class I evidence is often defined in a way similar to “level I” above. An RCT became the gold standard for medical evidence, regardless of the type of question asked. The many biases and limitations of RCTs were disregarded. Moreover, often the important caveat that the study must be “properly designed” is not emphasized.

The problem is particularly severe in surgical fields. Surgical RCTs can be successfully performed on identical experimental animals under tightly controlled conditions. Applying this design to human subjects is vastly more difficult, and may be impractical in most cases. Certain basic problems are inescapable, because human disease is more variable than animal models, and most patients have additional illnesses and treatments that complicate the picture. Usually, patients have their own ideas about the treatment they want; needless to say, so do surgeons – so
true randomization is nearly impossible. The definition of the study, alternative, and control treatments is very difficult to agree upon and is often deficient. Really blinded design involves sham surgery and is rarely feasible, so new treatment effects and placebo effects abound. The surgical treatment itself is variable and hard to standardize. Operative plans and details as well as surgical skill vary widely, which must affect average outcome. The problem is that a procedure under the same name may be performed differently with different effects depending on the surgeon. Averaging the results between a better and a worse way to perform a procedure will not really help direct patient care. Surgical treatments are expensive, may involve implantation of costly hardware, so financial pressures abound – whether from industry trying to promote a new product, or government agencies trying to control costs. There is no agreement on the best way to analyze surgical study results. For example analysis by “intent to treat” is considered standard for RCTs; it is very useful in studies of medical treatment, helping to account for differences in compliance among the different treatment branches. Intent-to-treat analysis runs into problems in a study that compares medical and surgical treatments. If a patient crosses over from failing medical treatment to the surgical arm, the effect of surgical intervention is assigned to nonoperative management; the failure of medical treatment among patients who were assigned to surgery but refused it accrues to the surgical branch of randomization. Since surgery is a single irreversible step, comparison with nonoperative treatment is inherently asymmetrical.

The above list of problems is not news to clinical researchers. The design of surgical RCTs typically tries to address at least some of the inherent problems. Unfortunately the result is a new list of methodological issues. The reader of almost any published surgical RCT will usually be able to discern at least one fatal flaw:

1. Variability is addressed by carefully selecting patients, which often eliminates the great majority of patients from participating practices from the study. The population is too narrowly defined, so the conclusions cannot be generalized.

2. The problem of surgical indications – in order to randomize the patients ethically, equipoise must exist in the mind of the surgeon about the two options that are being compared. This is a rare situation, given the emotional connection most surgeons feel to their work and their patients. Equipoise must be established before starting the study to avoid bias and obtain uniform criteria for enrollment. Most studies so far have not addressed equipoise, though it is in principle feasible. Inevitably, limiting randomization to situations in which equipoise exists will further narrow the population of patients who are eligible to be study subjects.

3. Equipoise must also exist in the mind of patients, and the majority of the subjects must not change their mind after the study has begun. Many studies suffer from excessive crossover.

4. The study surgeons do not represent the population of surgeons likely to perform the procedure – either inexperienced surgeons are allowed to participate, skewing results one way, or the study is excessively centered on a few major academic centers, resulting in opposite bias.

5. A systematic bias in the study design, determined by the funding source. For example, a study funded by a government agency, where spine surgery is offered to an ill-defined population of patients who may not need it, and the conclusion is that spine surgery is not effective.

6. The population is defined is such a way as to exclude patients who may be helped most. For example, the design insists on performing a particular type of imaging that takes a long time to arrange and perform, delaying care beyond the time frame in which it is most likely to help.

7. Results depend on the selection of an artificial cutoff point for continuously variable parameters like aneurysm size, or amount of spondylolisthesis. This is a well-known statistical fallacy.

One example of a prospective randomized study that is more misleading than helpful is a quite typical study funded by the Medical Research Council (MRC) in Britain and published in 2008. In that study patients with back pain were randomized to receive stabilization surgery or intensive physical therapy but only in cases where both patients and their surgeons were not sure surgery was needed. There were no imaging criteria and the procedures were not standardized. A total of 349 patients were recruited (176 surgery and 173 rehab) in 15 participating centers, and followed for 2 years. Results of the study were reported as follows: 81% follow-up at 2 years; crossover: 28% of patients from rehab group had surgery at the end of 2 years of follow-up. On the other hand, 7% of those randomized to have surgery opted for rehab. By intent to treat analysis there was a small but statistically significant effect of surgery in improving ODI (-4.1). No difference was seen in any other outcome measure. The authors concluded that there was no clear evidence for the benefit of surgery over rehabilitation in treatment of chronic lower back pain.

There was no consistent clinical, imaging or biomechanical definition of the need for surgical intervention. Randomizing patients who were not perceived by their surgeons as having a clear indication for surgery resulted in including patients who may have not been offered surgery under normal circumstances. We only offer surgery when we are sure it is indicated. The crossover was large and asymmetrical with results of surgery accruing to 28% of the rehab group in the intent to treat analysis. The number of patients appears impressive, but 349 patients must
have been a very small part of the total spine practice in 15 major centers. The strong conclusion was not justified by the data and thus can only mislead. The data supported only the following, more modest conclusion: “When the surgeon and the patient are not convinced that an operation is necessary for back pain, and without clear imaging criteria, in the non-representative population that was studied, ill-defined surgical treatment appeared to offer little benefit over physical therapy; provided you counted the outcome of surgery in those who had an operation after being assigned to physical therapy and failing it, as outcome of physical therapy.”

Creating the empirical classes of evidence led to a tyranny of RCTs. These days RCTs are promoted as the key source of data for decisions making; they are also a great tool of career advancement. The irony is that “level I evidence” was defined as that which comes from “properly designed” randomized controlled studies. Since almost all surgical RCTs suffer from profound methodological flaws, they do not amount to level I evidence. The prevalence of severe errors renders many surgical RCTs controversial at best. The drive to obtain “class I” status results in massive financial investment in these studies, often using taxpayer money. The large investment creates its own source of bias. The acceptance of deeply flawed results in leading journals provides misleading “punch lines” to the general public and the rest of the medical profession. Worse, because of the prestige associated with large multicenter trials, the flawed conclusions may influence practice and cause actual harm. A fatally flawed RCT is worse than no data at all.

So, am I against evidence-based medicine? Of course not— it is good to base clinical practice on scientific evidence when good evidence is available and relevant to your clinical question. I am just against the tyranny of flawed RCTs. Before an RCT or any larger scale trial can be planned, a lot of data have to be accumulated some other way, in order to be able to generate testable hypotheses. Then, when a medical hypothesis is generated, the best study design to test it can be chosen. In many cases the best design for that particular question will not be an RCT. For example, the answer to the important hypothesis “The rate of complications of operation X is lower than the rate of success” will be provided by an objective, unselected, and prospective data collection, but not an RCT. Excessive emphasis on RCTs may have the effect of stifling medical research by devaluing the type of work that helps generate hypotheses. An excellent, well-documented case report or a short series can raise questions and propose innovative approaches. Just think about the impact of a single “anecdotal” but well-documented report of cure of a glioblastoma.

Randomized controlled trials will have their place in neurosurgery. They have to be designed very carefully and avoided when a good design turns out to be impractical or not feasible. It is sometimes possible to design a very good randomized, sham-controlled trial in neurosurgery. One such trial was the 2003 Olanow et al. study on fetal transplantation for Parkinson’s disease. The article was critically important, profoundly affected the field of functional neurosurgery, and remains an example of true level I evidence. The conditions that allowed a true-blinded RCT design in that case do not exist in the vast majority of clinical situations in neurosurgery. The current efforts to establish surgical equipoise before randomization, like the recent work by Ghogawala et al., are also commendable.

Inclusive, blinded, and audited databases with long term follow-up and objective outcome scales can be very useful. Such a cohort study, although it is not called “class I,” has less potential for bias than an RCT. Although cohort studies are not specifically designed for comparisons between treatments, some comparisons between clinical choices and practice styles can be cautiously made. It does not answer the same questions an RCT can, but is also less likely to mislead. I would support revision of the nomenclature, avoiding assignment of levels or classes of evidence excessively reliant on intended study design, and creating a more nuanced system of classification of medical science, based on suitability of the design to the research question asked, the ability of the study to affect actual practice, and the ability of the researchers to maintain quality within the intended study design.

Overturning the tyranny of RCTs will also save money, since a poor RCT still costs a fortune. Public funds may be better used if diverted from RCTs to more investment in objective audited surgical outcome databases.

REFERENCES

1. Ghogawala Z, Martin B, Benzil EC, Dziura J, Magge SN, Abbed KM, et al. Comparative effectiveness of ventral vs dorsal surgery for cervical spondylotic myelopathy. Neurosurgery 2011;68:622-30; discussion 630-1.
2. Olanow CW, Goetz CG, Kordower JH, Stoessl AJ, Sossi V, Brin MF, et al. A double-blind controlled trial of bilateral fetal nigral transplantation in Parkinson’s disease. Ann Neurol 2003;54:403-14.
3. U.S. Preventive Services Task Force. Guide to clinical preventive services: Report of the U.S. Preventive Services Task Force. DIANE Publishing; 1989. p. 24.
4. Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Blood E, Hanscom B, et al.; SPORT Investigators. Surgical versus nonsurgical therapy for lumbar spinal stenosis. N Engl J Med 2008;358:794-810.
5. Wilson-MacDonald J, Fairbank J, Frost H, Yu LM, Barker K, Collins R, et al.; for the Spine Stabilisation Trial Group. Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation programme for patients with chronic low back pain: The MRC spine stabilisation trial. Spine (Phila Pa 1976) 2008;33:2334-40.