
Where Is the Evidence? Best Practice in Spine Surgery

5

Wim Van Biesen

Contents

5.1	Introduction	77
5.1.1	Why Evidence-Based Medicine and Guidelines: A Short History	78
5.1.2	Are Systematic Reviews, Guidelines, and Clinical Performance Measures Birds of a Feather?	80
5.1.3	Problems of Evidence-Based Medicine	83
5.2	What About the Patient?	85
5.3	Conclusion	86
	References	88

5.1 Introduction

Evidence-based medicine (EBM) is gaining importance in our current health-care landscape. However, the concept of EBM risks to become a victim of its own success if all parties involved are not clear on what EBM really is, why and how EBM should (not) be practiced, and have sufficient skills to distinguish methodologically sound papers from biased opinion papers.

We will discuss that improving patient outcome does not only require attention to high-quality evidence but also understanding of the processes of medical decision-making. We will advocate that rigorous methodology is the cornerstone of guideline production, but in those cases where quality evidence is not retrieved, consensus-based guidance might be suitable to assist the practicing spine surgeon at the bedside. We will advocate that EBM should rather aim for transparency than for statements carved in stone. Last, but certainly not the least, we will argue why and how EBM should be supportive for involving the patient in a shared decision-making process.

It should be well understood that the more the effect of certain interventions is, (1) consistent and accurately predictable, (2) clinically relevant to patients rather than affecting surrogate outcomes, and (3) a priority for patients and other stakeholders, the more likely it is that adherence to the provided guidance will improve

W. Van Biesen, MD, PhD
Department of Renal Division, Ghent University
Hospital, Ghent, Belgium
e-mail: Wim.VanBiesen@UGent.be

the outcome of patients, and the more desirable it is that the final result of the shared decision-making process is in line with the provided guidance [1, 11].

This necessitates special attention to how guidance is provided and to understand medical decision-making processes.

5.1.1 Why Evidence-Based Medicine and Guidelines: A Short History

All physicians, including spine surgeons, want to give the best treatment to their patients. All patients want to receive the best treatment from their physicians. All health-care providers want their physicians to provide the best treatment to their patients. So far so good, but what does it mean to be “the best treatment” and how do we find it?

In the past, all available medical knowledge could easily be assimilated in one person. In addition, for most of the topics covered, there was a direct and concrete relation between the intervention performed and the outcome observed, for example, the use of antibiotics in infections or of sterile procedures in surgery. This made it relatively easy for physicians who treated a reasonable number of patients to learn from experience what to do in which situation. Over the last decades, we have seen an exponential growth of available “evidence” of varying quality, resulting in much noise, but only a limited signal. Whereas everybody can nowadays access all this information online, there is so much information available that no single individual can digest it. It looks like the thinking is global but the treatment local.

In addition, we are now dealing with improvements and outcomes that are only observable after a long follow-up time and in larger populations (e.g., decrease of cardiovascular risk by the use of statins). All these developments created the necessity to have experts summarizing, interpreting, and translating the available information. We should be largely informed to know which knowledge we can delete and which can be used

to become a better spine surgeon. This resulted in the conception of evidence-based medicine (EBM) [2, 3]. The concept of EBM is that all medical actions should be backed up by a thorough and systematic search for and analysis of the available evidence. Therefore, a systematic methodology, mostly denominated with the acronym PICOM (Table 5.1), was developed to search for the available evidence in what is now called *systematic reviews*. PICOM aims to correctly identify the different components of the search: patient, intervention, comparator, outcome, and methodology. The underlying idea was that as for regular scientific experiments, everybody should be able to redo the search and come up with the same papers and evidence.

PICOM focuses on “patients,” to assure external validity of the studies that one is searching for: are patients in this study comparable to the ones I see in clinical practice? Many RCTs have quite stringent in and exclusion criteria, which might result in the fact that the study population is not at all representative for a patient with this condition [4]. Also a good description of the “intervention” and “comparator” is of importance. It should be checked whether these are relevant, reasonable, and in line with expected practice. Often neglected is the “outcome” definition. This is critical, as it will determine whether or not the outcome is relevant to the patient and whether the effect size is really meaningful. Many studies tend to report surrogate outcomes in qualitative ways (better, improved, etc.), which does not clarify *what* exactly was improved by *how much* of a certain measure. It is best to look for hard rather than surrogate endpoints and to report them as absolute effect size rather than as relative changes. In contrast with “narrative reviews” where authors base their interpretations on the literature available to them from their own experience (also known as eminence-based medicine), a systematic review using the PICO methodology makes sure that *all* available evidence will be retrieved for analysis and not only the evidence that fits the ideas of the authors.

A lot of progress has already been made in the methodology of how to compile and extract evidence [5–7]. Whereas a rigorous methodol-

Table 5.1 PICO methodology to support systematic searches

Nonspecific clinical question	Question type	PICO	Specific question
What is the best intervention for a cervical hernia?	Intervention	P: adult patients with a proven cervical disk herniation C5–C6 I: surgical treatment with method A C: surgical treatment with method B O: reduction of pain as measured with VAS, restoration of function M: randomized controlled trial	In adult patients with a proven cervical hernia C5–C6, is surgical treatment with method A as compared to method B superior to reduce pain as measured by VAS and to restore functional capacity?
What is the best method to diagnose a cervical hernia?	Diagnostic	P: adult patients with symptoms compatible with cervical disk herniation I: magnetic resonance imaging (MRI) C: computed tomography (CT scan) O: diagnosis of cervical disk herniation M: cross-sectional design	In adult patients with symptoms compatible with cervical disk herniation, is MRI more specific and sensitive as compared to CT scan to diagnose a cervical disk herniation?
Does smoking increase the risk for cervical disk herniation?	Etiology	P: adult patients I: smoking C: nonsmoking O: cervical disk herniation M: cohort study	In adult patients, does smoking versus nonsmoking increase the risk to develop a cervical disk herniation?
What is the risk of developing a relapse after the resection of an intraspinal meningioma?	Prognosis	P: adults I: patients who have been successfully operated of a benign intraspinal meningioma C: healthy persons of the same age O: symptomatic relapse of meningioma M: cohort study	In adult patients who have successfully been operated from a benign intraspinal meningioma, as compared to healthy age-matched persons, what is the relative risk of developing a symptomatic relapse?

We distinguish four study types: interventional trials, diagnostic trials, etiological trials, and prognostic trials. Each of these study types has a preferred study design to answer the question

For a diagnostic trial, there is a need for a golden standard as a reference to calculate sensitivity, specificity, and positive and negative predictive value of a test. When such a golden standard is lacking, one can convert a diagnostic study to an interventional trial by linking it to an intervention: e.g., does surgical treatment of a cervical disk herniation as diagnosed by MRI versus EMG versus CT scan lead to better outcomes as defined by a better functional recovery
P patient population, *I* intervention, *C* comparator, *O* outcome, *M* preferred methodological design

ogy is an indispensable step in providing high-quality guidance, the low availability of quality evidence for many areas in medicine remains a major hurdle. It appears that for many conditions in specific populations, there is insufficient evidence to meaningfully support a statement. Hardcore EBM adepts believe that in these conditions, no conclusions can be made. However, this would leave the clinician without guidance for many topics as there is a scarcity of high-quality clinical trials (especially in the domain of spine surgery) and absence of high-quality evidence for many conditions. In these circumstances, a

compilation of additional expertise can help clinicians out in daily clinical management. However, it should be made *transparent* that in such cases the guidance is based on consensus rather than on evidence and be made clear that a systematic literature search has confirmed the absence of firm evidence: rigorous methodology should always be the starting point. There is thus a big difference between *believing* there is no evidence and *concluding* after a systematic search that there is no evidence. The use of PICO methodology for a systematic search and presentation of the findings in objective, easy-to-read data extraction

tables providing both the evidence and the quality assessment of the evidence is one promising way to achieve this goal [8, 9]. In addition, tools like the AMSTAR scoring system [10] can help the clinician to assess the methodological quality of a systematic review (Table 5.2). Good quality systematic reviews, such as those performed by the Cochrane Collaboration, should be the cornerstone for all medical interventions and patient care for the individual health-care worker. The strength of the available evidence can also be formally assessed, e.g., by using a system such as GRADE. Within GRADE, studies start from a certain level of strength, depending on the type of studies available (see Chap. 4). As a standard, randomized controlled trials are considered higher quality evidence than observational studies or case reports. However, strength of evidence can decrease or increase following evaluation of prespecified criteria. Well-performed large observational trials with low risk of bias can thus score higher than badly performed RCTs at high risk of bias. Within GRADE, the level of evidence is qualified as A, B, C, or D, where A stands for high quality where additional evidence is unlikely to change the conclusion [1].

5.1.2 Are Systematic Reviews, Guidelines, and Clinical Performance Measures Birds of a Feather?

Whereas systematic reviews are suitable to answer individual, well-defined questions, there is a need for organizations that interpret the available evidence at a broader societal level in a transparent and methodologically robust manner and provide guidance on how certain conditions should be managed. This is the best guarantee for maintaining sustainable and fair health care. Guidelines can protect physicians from prescribing treatments that are ineffective, where the term “ineffective” covers different meanings ranging from “not working at all” to “not improving relevant outcomes” or “achieving outcomes that are not a priority” [11]. Guidelines can be used to steer health-care policy, as is already incorporated

in the GRADE system [12], and can serve to decrease the pressure of industry or public opinion to prescribe ineffective interventions. The GRADE system explicitly states the level of recommendation as “strong” (we recommend) or “weak” (we suggest), where it is critical that this appraisal can be completely independent from the evaluation of the strength of the evidence. The strength of recommendation only depends upon a judgment of the desirability of the recommended action [1]. Guidelines differ thus from systematic reviews, although ideally, they should be based on them, as they also incorporate “value” attributed to certain outcomes and not to others, and thus indirectly allow the necessary prioritization to build up health-care strategies.

Some may argue that guidelines will be (ab)used by payers and policymakers to monitor and judge the quality of care provided by physicians. There is plenty evidence that this should be done with utmost care. In the first place, all partners involved should be aware of what exactly is being measured and why (Table 5.3). Quality assessment is especially dangerous when based on indicators that not only reflect center performance but also individual patient preferences, e.g., the choice for a certain type of intervention versus another [13]. Instead, it is preferable to focus on developing indicators that reflect the extent to which units facilitate shared decision-making in certain fields by offering alternative treatment options yes or no. In this respect, spine surgeons should work in a multidisciplinary team, including colleagues of the pain clinic. Furthermore, the choice of which indicators one will use to monitor clinical performance may heavily affect the clinical result that clinicians will aim for in reality [11]. If a level “X” is claimed to be the best value for hemoglobin, should one then aim to have the mean of the population at “X,” meaning a substantial part is below “X,” or should one measure the percentages of patients above “X,” resulting in many patients far above “X,” which might be undesirable. When applying indicators based on percentages of patients that achieve a given target, one should be certain that the preset percentages are achievable in clinical reality [14] without jeopardizing “personal choice” [15] or without inducing “cherry picking.”

Table 5.2 AMSTAR score for evaluation of quality of systematic reviews

Evaluation point	What is evaluated?	Why it is of importance?	What should you look for?
1. Was an "a priori" design provided?	The research question and inclusion criteria should be established before the conduct of the review	Make sure that it is clear what type of evidence is looked for and on which criteria studies will be in or excluded from the analysis. It reduces the risk that only studies with "desired" results are selected	Authors need to refer to a protocol, ethics approval, or predetermined/a priori published research objectives to score a "yes"
2. Was there duplicate study selection and data extraction?	There should be at least two independent data extractors and a consensus procedure for disagreements should be in place	Two people know more than one; avoid (un) (intentional) mistakes in the data extraction and biased interpretations	Authors should mention who has done study selection and who has done the data extraction. It should be specified how disagreements between the reviewers were resolved
3. Was a comprehensive literature search performed?	At least two electronic sources should be searched. The report must include years and databases used (e.g., Central, EMBASE, and MEDLINE). Keywords and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study and by reviewing the references in the studies found	The reader should be able to repeat the search and come up with the same results. The search should be as exhaustive as possible to be sure all available evidence is discovered	Authors should at least mention two databases they have searched, and their search building strategy should be available
4. Was the status of publication (i.e., gray literature) used as an inclusion criterion?	The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports from the systematic review, based on their publication status, language, etc.	To avoid publication bias, it is important that also studies that are not published in official peer-review papers or in English are retrieved, e.g., conference proceedings, abstracts, and also trials that have been done but have never been analyzed	Authors should state they searched for gray literature or unpublished literature
5. Was a list of studies (included and excluded) provided?	A list of included and excluded studies should be provided. In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions, and outcomes. The ranges of characteristics in all the studies analyzed, e.g., age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases, should be reported	For all included studies, the basic characteristics of the population included should be clear, what exactly were the intervention and the comparator and what the observed outcomes were. This is necessary to allow a balanced appreciation of the available evidence and to verify whether the studied population is comparable to the population you are treating	The characteristics of the included studies must be listed in a table or other systematic form of presentation. For excluded studies, it can suffice to mention the reason why they were excluded and a reference
6. Were the characteristics of the included studies provided?			

(continued)

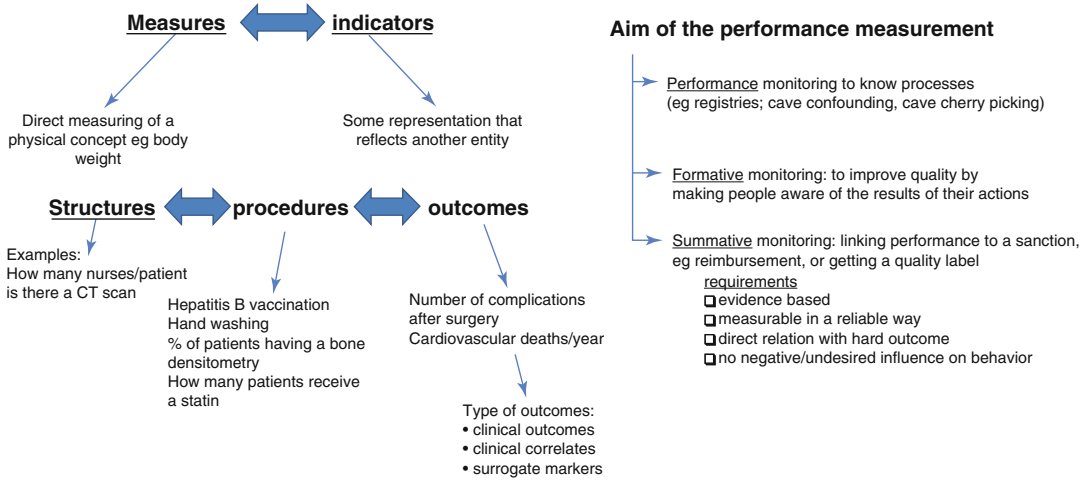
Table 5.2 (continued)

Evaluation point	What is evaluated?	Why it is of importance?	What should you look for?
7. Was the scientific quality of the included studies assessed and documented?	“A priori” methods of assessment should be provided (e.g., for effectiveness studies if the author(s) chose to include only randomized, double-blind, placebo-controlled studies or allocation concealment as inclusion criteria); for other types of studies, alternative items will be relevant	This question explores whether the authors systematically evaluated and presented the different forms of risk of bias in the individual studies, e.g., using the Cochrane risk of bias scale	Authors should mention how they assessed the risk of bias of the individual studies. The results of this risk of bias analysis should be presented
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review and explicitly stated in formulating recommendations	To avoid that strong conclusions are based on studies with high risk of bias	There should be a reference to the presented evidence and its quality for the majority of the major conclusions of the systematic review
9. Were the methods used to combine the findings of studies appropriate?	For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e., chi-squared test for homogeneity, I ²). If heterogeneity exists, a random effect model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e., is it sensible to combine?)	Papers can only be interpreted together if they handle the same topic and the same intervention and are performed in a comparable population, and this is to avoid that apples are compared to oranges to come to a conclusion about bananas. This test evaluates how well the individual studies handle comparable situations. If they do not, they should not be lumped together	Authors should mention they evaluated heterogeneity and present results
10. Was the likelihood of publication bias assessed?	An assessment of publication bias should include a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test, Hedges-Olken)	Publication bias can be problematic, as smaller, negative studies have a high probability to be not published	Look for terms as “publication bias”
11. Was the conflict of interest included?	Potential sources of support should be clearly acknowledged in both the systematic review and the included studies	Conflict of interest can lead to biased presentation and/or interpretation of the results. Whereas financial conflict of interest is most obvious, one should not neglect the impact of intellectual conflict of interest	Are the affiliations of all authors clarified for both the systematic review as for the individual studies? Are any potential financial interests of the authors mentioned?

Adapted from Shea et al. [10]

For each question, the answer is either yes, no, or cannot answer. Every “yes” provides one point. The maximum score is 11

Table 5.3 Quality performance measures and indicators



In clinical performance measurement (CPM), one should distinguish between *measures* and *indicators*. By definition, quality is not a physical property, so it can only be *estimated* by *indicators*. As a consequence, quality is always estimated *indirectly* by a surrogate marker. Evaluation of quality can be done at 3 different levels.

At the level of “structural “ items, we can assess if some or other *logistical* item is yes or no present. One should be aware that structural items are easy to achieve (though they might be expensive), as mostly, they can be purchased in some way or the other. However, it should be clear that structural items could be a *necessary* item for quality, but that they never are *sufficient* in themselves, as assessing structural items doesn’t tell anything about how they are implemented in practice.

Assessment at the level of “*procedures*” evaluates how organizational items are implemented in clinical everyday practice. As such, they reflect how good the adherence is to recommended procedures. However, this indicator can be confounded by cherry picking, and risks to ignore the process of shared decision-making.

Assessment of outcomes seems to be the most straightforward way to evaluate clinical performance; however, this is prone to cherry picking, and might lead to unwarranted practices if this assessment is coupled to summative monitoring but based on wrong indicators. Summative monitoring uses CPM to link performance to some form of reward (mostly reimbursement) or punishment. It is only acceptable if the indicator used is a valid reflection of an evidence based practice and when implementing the indicator does not result in omission of the shared decision making process.

Whereas it seems logical to use performance indicators for which we have a solid evidence base and which are considered a priority [16], clinical performance measurement (CPM) initiatives tend to select their performance indicators based on feasibility, implying that some (potentially more) important aspects of care may be neglected due to the mere fact that they are presumed to be “difficult to measure.” For example, in the CPM project of KDOQI [17], 36 of 114 guideline recommendations were originally identified as having a high priority. However, 14 of these recommendations were not transformed into performance indicators partly because they could not be unambiguously made operational for measurement purposes.

5.1.3 Problems of Evidence-Based Medicine

Many believe that the highest degree of evidence is coming from randomized controlled trials (RCTs). However, RCTs only provide evidence if they are free from bias, whereby bias should be understood as any process that systematically causes the true effect to be different from the observed effect. Some forms of bias are obvious and well described, e.g., failure to adequately blind the intervention and the comparator (Table 5.4). Several different scoring systems have been developed to search for and quantify presence of bias in randomized controlled trials, e.g., the Cochrane risk of bias tool [18]. However, some other forms of bias are not as explicit and

Table 5.4 Risk of bias in randomized controlled trials

Type of bias	What is it?	How to look for it?	Consequence if not done properly
Sequence generation	Was the order of inclusion in the intervention versus comparator groups well performed?	Do authors refer to a random generator table or use of random generator software? Coin tossing, throwing dice, and shuffling sealed unmarked envelopes are also acceptable. Randomization methods based on parameters that may be associated with the outcome should be avoided	Differences between intervention and comparator group might not be random
Alllocation concealment	Can any person involved in the management of the patient know in which group the patient will end up?	Central allocation (telephone, noninvolved pharmacy, etc.); use of sequential sealed opaque envelopes	Patients might be allocated to groups based on certain criteria other than random, which might create differences between groups
Blinding	None of the involved persons know which treatment the patient is on Double blind: neither the patient nor the physician know which treatment is on	OK if: Assessor of outcome is blinded Patient and physician are blinded, and note that blinding is only important for subjective outcomes and less for hard outcomes or subjective outcomes assessed by a third independent (blinded) person	Unblinding might result in effect sizes bigger than reality, as one can overestimate the effect
Incomplete/selective outcome reporting	All prespecified outcomes are reported and this is in the prespecified way, using the prespecified measures and presentation	Prespecified outcomes are not reported Outcomes are not reported for all patients randomized or assessed for study inclusion As treated analysis reported only Non-prespecified outcomes are reported Inappropriate missing data imputation techniques No mentioning of adverse events	Only outcomes that are in the desired direction are reported, and those in conflict with the original hypothesis are ignored Effect size is overestimated; adverse effects are underestimated Reporting non-prespecified outcomes results in noise generation
Risk of sponsor bias	The study was sponsored by industry	Study design such that it is beforehand obvious that the study drug will be better than comparator (e.g., by incorrect dosing of the comparator) Pointless studies (me too study, use of invalid surrogate markers) Incorrect or biased interpretation of the results	Studies sponsored by industry tend to have a higher likelihood of having results that favor the study drug
Selection bias	The study population is not representative for the population the intervention is aimed at	Check inclusion and exclusion criteria Check whether demographic data of the study population are comparable to those of the population of interest	Selection bias might hamper external validity of the study

can even be only apparent to those who have hands-on experience with the intervention or comparator or require in-depth knowledge of statistical techniques and epidemiology. For example, using an as-treated versus an intention-to-treat analysis can lead to different results and different conclusions. Last, RCTs should not only provide valid (methodologically correct) but also relevant (does it matter?) results. One can distinguish hard endpoints, i.e., endpoints that matter directly to patients, such as death, quality of life, loss of vision, etc., and surrogate endpoints. These surrogate endpoints are mostly parameters that do not directly matter to the patient. Such a surrogate parameter can be a valid representation of a hard endpoint, but this is seldom the case. The reason why many RCTs opt to have surrogate rather than hard endpoints is that surrogate endpoints are more easy to be accrued and that they can be mostly observed after a short observation period, whereas hard endpoints mostly take a certain duration of time to occur. However, plenty examples are available in the literature demonstrating that interventions that improve surrogate outcomes do not result in improvement of the hard endpoints. Therefore, one should avoid the use of surrogate endpoints. Besides the outcome itself, also the size of the observed effect is of importance: sometimes a statistically significant effect can be completely meaningless from a clinical perspective. Therefore, it is dangerous to use quantitative expressions (better, improved, etc.) rather than absolute effect sizes (expressed in numbers how much difference was obtained with the intervention vs. the comparator).

Whereas high-quality RCTs reporting hard endpoints are scant, those that are available make us realize another challenge: this type of studies is expensive and mostly performed by pharmaceutical companies [19, 20]. This results in the Catch-22 situation that evidence is mostly created, and thus only available, for (often expensive) newer drugs or interventions. Very rarely funding is available to investigate cheaper alternatives; so where two sides of the coin should be evaluated, the cheaper one remains invisible. As a consequence, public funding bodies should support research for alternative (cheaper) treatments

that will not receive support from industry, while investigators should also be open for ways to run these studies at lower costs. Systematic searches can make visible where evidence is lacking, or bias, and confounding is apparent and where further studies are needed. In the field of surgery, an additional problem is that the outcome of a surgical intervention is also sensitive to differences in experience, that interventions are difficult to standardize, and that no placebo effect can be organized. As a consequence, the success of two different treatments can be different for two different surgical teams.

5.2 What About the Patient?

Often the most cumbersome aspect of EBM-based guidance regards the actual implementation after release [21]. Guidance-producing bodies should aim to generate guidance that supports shared decision-making [22]. This process goes far beyond simply explaining different options to the patient [23] and is complicated by the existing uncertainty [24]. Of note, shared decision-making concerns two relationships: one between the treating physician and the patient and a second between the guidance-producing body and its users. Guidance should be formulated in a way that it allows to make a decision based on the provided evidence. This is a change in paradigm from “we will tell you what is best” to “we will provide you with the data in order to come to your own conclusion.” As already stated, when there is more convincing evidence, it becomes more likely the decision-maker will opt to come to the same conclusion as the guidance-producing body, an aspect that is very well captured in the GRADE system [25]. One should be aware that the nomenclature for rating guideline recommendations is complex and thus inhibits the knowledge dissemination process. The two-step rating of the GRADE system [1, 25] into *strength of recommendation* (level 1, 2, or not graded) and *quality of the supporting evidence* (A, B, C, or D) is often neglected. This can result in a misinterpretation of the guideline, as the implications for patients, clinicians, and policy

are not considered and the quality of the available underpinning is neglected.

Whereas we all believe in the value of “objective” information, it has been well recognized that human beings do not take decisions solely on rational grounds [26]. On the contrary, emotional grounds often will play an important role in the shaping of a preference. This process can be guided by rational argumentation, be it only partly [27], as personality traits will determine which arguments will result in which emotion and finally which decision will be made. People value much more consequences that are nearby in the future than remote ones, and they also commonly prefer “avoiding harm” over “creating potential benefit” [28]. Patients will, e.g., be easily compliant with their prescribed medication to avoid discomfort, e.g., itching (direct harm), but it will be more difficult to convince them to take drugs that avoid cardiovascular events in the longer term (remote benefit) or to improve a surrogate marker such as HbA1C. Similarly, also physicians are not free from “inborn thinking errors,” such as *anchoring* (being fixed on one specific interpretation of the evidence and ignoring other potential maybe even more plausible pathways), *attribution* (linking two consecutive events to each other in a causal way, e.g., I took a pill and the cough disappeared, so the pill made the cough go away), and *availability* (being influenced in a medical decision by a previous dramatic or unusual course of a certain disease, often linked with the desire to avoid harm and which might lead physicians to take unnecessary precautions or avoid beneficial treatments because they have very rare but serious side effects). Providing evidence in a structured manner on the risks and benefits of alternative treatments might help to avoid these thinking errors [29]. Making physicians and patients aware of all available options can, e.g., avoid anchoring effects. Considering the influence of anticipated emotions in the decision-making process, spine surgeons should rather be supported to “predict” outcomes of a certain intervention and explain the potential consequences of taking decision A versus B. In that

way, the patient himself can make the decision, taking into account his own preferences.

Presenting the available evidence in a format that supports the balancing of pros and cons of all potential interventions will most likely lead to outcomes that are preferred by the patient. In fact, from the very beginning, the framing of the question to the needs of the individual patient was incorporated in the early definitions of evidence-based medicine [3].

5.3 Conclusion

The universal “best treatment” does not exist. Evidence-based medicine is more than a fixed, rigid way to use statistical inference for the assessment of available evidence: it is a way of performing medicine with a critical mind for all steps involved in the care of the patient [9]. We believe that the best way to assure sustainable improvement in patient outcomes is to support shared decision-making between spine surgeons and patients by providing the best available and well-balanced evidence in a format that allows the patient (and sometimes also the physician) to see all pro’s and con’s of the available options in an understandable way.

Editor’s Note on Evidence

As described in this and the previous chapter, evidence-based medicine (EBM) is gaining importance in our current health-care landscape. The scope of those who financially contribute to the health-care systems (the tax payers, you and me) and of those who make the decisions on priority spending (politicians and insurance companies) changes from effective to efficient, efficient in the sense of cost-effectiveness. Automatically, if effectiveness is the denominator and cost is the nominator, EBM comes into view. Although it is hard to calculate the cost of a given treatment and its economic gain, determining its effectiveness is even harder to accomplish.

This is especially true for all kinds of surgical treatments but for spine surgery in particular. It is not evident to demonstrate evidence in spine surgery for several reasons:

- First, there is no such thing as a “sham” surgical procedure. This makes comparative studies difficult. An RCT conducted about medication can always compare with placebo treatment. In spine surgery, there is no placebo control.
- Second, spine surgery is mostly offered in case of failed conservative treatment. As such, an RCT between “natural evolution” and conservative and surgical treatment is difficult to realize. In most attempts of this kind, the crossover from the conservative treatment group to the surgical one makes it hard to draw significant conclusions.
- Third, in spine surgery, the most important outcome measurement is pain; nothing is more difficult to measure, because so influenced by all possible known and unknown factors. Because pain is a very subjective parameter, we often look for more objective parameters to measure, such as radiological outcomes. Unfortunately, the correlation between good radiological outcome, such as fusion after arthrodesis, and good clinical result is rather poor.
- Finally, in most clinical studies, good clinical outcome is reported after spine surgery. I am not aware of any study where the outcome after spine surgery is very bad. Probably, we only conduct clinical studies once we are convinced that the outcome must be good. If our null hypothesis risks to be rejected, we often adapt our study protocol or material and methods in a way we can accept our hypothesis that a given treatment (spine surgery) is beneficial. Sometimes

we study conditions with a good natural outcome anyway (lumbar disk herniation, Modic type 1 changes), even if “blurred” with some kind of surgery. Often, the least invasive surgery does alter the natural evolution in the slightest way, offering good clinical results.

If clinical outcomes after spine surgery are generally good, it becomes hard to compare two different surgical treatment options. If we want to demonstrate a significant difference between two good treatment options, the number of participants in the two study groups must be enormous to have enough power. It takes a lot of money to conduct such an RCT. In most cases, the industry is the only partner that can provide this necessary funding. It is reasonable to accept that they will only be willing to do so, when the chance that patients treated with their implants do better than the control group is high. In case they really have better results, everybody will disregard this outcome due to “sponsoring bias.”

Nevertheless, EBM should be the cornerstone for our spine surgery research. By accepting this scientific methodology, we should rather aim for transparency of what we do, rather than for statements carved in stone. Good science should be vulnerable for critical evaluation and for new empiric data. As such, it is an on-going process. What is true today, changes tomorrow. However, it is not because EBM is lacking or because good science is an on-going process that the individual spine surgeon can do whatever he/she thinks is best. Every decision prior to spine surgery should be the result of a shared decision-making process in which EBM, if present, should be supportive to that decision, and in case no EBM is available, internationally accepted guidelines should be respected. This seems normal and logic. But “logic” is, even for us spine surgeons, not always easy.

We are not well trained to think very rational. In reality, there is a gap between our thinking on statistics and our thinking on individual cases. Our brain likes causality. Statistical results with a causal interpretation do influence our brain more than when this causality is not present. It is generally accepted and scientifically recorded that EBM after RCT can change our attitude when dealing with a particular pathology (although this takes at least 2 years for most of us). But, even very convincing statistical results with a high degree of causality can seldom change personal beliefs, based on one or a few cases. We all like the “see one, do one, teach one” surgical strategy. Surprising results of individual cases do influence our decision-making more than the results of an RCT. Therefore, it requires some intellectual effort to accept the results of good scientific work and to give them more value in our daily decision-making, than personal experience.

Especially in spine surgery, rigorous methodology is the cornerstone of guideline production, but in case where quality evidence is not retrieved, consensus-based guidance might be suitable to assist us at the bedside. These guidelines should be as universal as possible. In the absence of evidence, these internationally accepted guidelines should be clear enough to reassure those who pay for what we do, that their money is well spend. This might seem obvious, but it is not. Besides a splendid description of surgical techniques, this book is an attempt to provide some guidelines on some topics.

In the next chapters, whenever appropriate, I provided an Editor’s note on evidence. Often, there is no evidence. In these cases, I tried to describe guidelines, realizing that, in most cases, internationally accepted guidelines for a specific spine pathology are lacking too.

References

1. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O’Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer T, Varonen H, Vist GE, Williams JW Jr, Zaza S (2004) Grading quality of evidence and strength of recommendations. *BMJ* 328:1490
2. Fouque D, Laville M, Haugh M, Boissel JP (1996) Systematic reviews and their roles in promoting evidence-based medicine in renal disease. *Nephrol Dial Transplant: Off Publ Eur Dial Trans Assoc Eur Renal Assoc* 11:2398–2401
3. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn’t. *BMJ* 312:71–72
4. Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, Shirlcliffe P, Aldington S, Beasley R (2007) External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax* 62:219–223
5. <http://www.thecochranelibrary.com>. Accessed Date Accessed
6. webpage E. <http://www.european-renal-best-practice.org>. Accessed Date Accessed
7. Nagler E (2013) ERBP Guideline development methodology: towards the best possible guidelines. *Nephrol Dial Transplant: Off Publ Eur Dial Trans Assoc Eur Renal Assoc*
8. Rosenbaum SE, Glenton C, Nylund HK, Oxman AD (2010) User testing and stakeholder feedback contributed to the development of understandable and useful summary of findings tables for cochrane reviews. *J Clin Epidemiol* 63:607–619
9. Greenhalgh T, Howick J, Maskrey N (2014) Evidence based medicine: a movement in crisis? *BMJ* 348:g3725
10. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, Ramsay T, Bai A, Shukla VK, Grimshaw JM (2007) External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLoS One* 2:e1350
11. Bogdan-Lovis E, Fleck L, Barry HC (2012) ‘It’s NOT FAIR! Or is it? The promise and the tyranny of evidence-based performance assessment. *Theor Med Bioeth* 33:293–311
12. Uhlig K, Macleod A, Craig J, Lau J, Levey AS, Levin A, Moist L, Steinberg E, Walker R, Wanner C, Lameire N, Eknayan G (2006) Grading evidence and recommendations for clinical practice guidelines in nephrology. A position statement from Kidney Disease: improving Global Outcomes (KDIGO). *Kidney Int* 70:2058–2065
13. Tangri N, Tighiouart H, Meyer KB, Miskulin DC (2011) Both patient and facility contribute to achieving the Centers for Medicare and Medicaid Services’

- pay-for-performance target for dialysis adequacy. *J Am Soc Nephrol: JASN* 22:2296–2302
14. Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW (2001) Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA* 285:2871–2879
 15. Howard AD, Howard RS, Goldstein SL, Meyer KB (2012) Fistula first breakthrough initiative (FFBI): lessons about arteriovenous fistula prevalence goals. *Am J Kidney Dis: Off J Natl Kidney Found*
 16. Kotter T, Blozik E, Scherer M (2012) Methods for the guideline-based development of quality indicators – a systematic review. *Implement Sci: IS* 7:21
 17. Sugarman JR, Frederick PR, Frankenfield DL, Owen WF Jr, McClellan WM (2003) Developing clinical performance measures based on the dialysis outcomes quality initiative clinical practice guidelines: process, outcomes, and implications. *Am J Kidney Dis: Off J Natl Kidney Found* 42:806–812
 18. http://ohg.cochrane.org/sites/ohg.cochrane.org/files/uploads/Riskof_bias_assessment_tool.pdf. Accessed Date Accessed
 19. Coyne DW (2007) Practice recommendations based on low, very low, and missing evidence. *Clin J Am Soc Nephrol: CJASN* 2:11–12
 20. Snowdon C, Elbourne DR, Garcia J, Campbell MK, Entwistle VA, Francis D, Grant AM, Knight RC, McDonald AM, Roberts I (2006) Financial considerations in the conduct of multi-centre randomised controlled trials: evidence from a qualitative study. *Trials* 7:34
 21. van der Veer SN, Tomson CR, Jager KJ, Van Biesen W. Bridging the gap between what we know and what we do in renal medicine: improving the implementability of the European Renal Best Practice guidelines. *Nephrol Dial Transplant: Off Publ Eur Dial Transplant Assoc Eur Renal Assoc*
 22. Haynes RB, Sackett DL, Gray JM, Cook DJ, Guyatt GH (1996) Transferring evidence from research into practice: 1. The role of clinical care research evidence in clinical decisions. *ACP J Club* 125:A14–A16
 23. Llewellyn-Thomas HA, Crump RT (2013) Decision support for patients: values clarification and preference elicitation. *Med Care Res Rev: MCRR* 70:50S–79S
 24. Smith QW, Street RL Jr, Volk RJ, Fordis M (2013) Differing levels of clinical evidence: exploring communication challenges in shared decision making. *Med Care Res Rev: MCRR* 70:3S–13S
 25. Balslem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH (2011) GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 64:401–406
 26. Kahneman D, Slovic P, Tversky A (1982) *Judgment under uncertainty: heuristics and biases*. Cambridge University Press, Cambridge
 27. Stanovich KE, West RF (2008) On the relative independence of thinking biases and cognitive ability. *J Pers Soc Psychol* 94:672–695
 28. Rabin W (2000) Risk aversion and expected utility theory: a calibration theorem. *Econometrica* 68:1281
 29. Herrington WG, Levy JB (2008) Metformin: effective and safe in renal disease? *Int Urol Nephrol* 40:411–417