ELSEVIER

The Spine Journal 19 (2019) 2–7

THE SPINE JOURNAL

Editorial

Show me the evidence: Dealing with bias in the medical literature

NASS 2018 presidential address

Daniel K. Resnick, MD, MS, 2018 NASS President

Professor and Vice Chairman of Neurosurgery, University of Wisconsin School of Medicine and Public Health, Madison, WI

Taking care of patients with spine related pain and disability is challenging. Diagnosis is an imprecise art and there are a multitude of competing treatment modalities for any given pain syndrome. Interventions may be expensive and dangerous, and measuring objective outcomes is difficult and time consuming. Those of us who accept the challenge of treating this patient population need not only to be armed with the best available evidence but also need to understand the limits of the available evidence when making treatment decisions. The purpose of this address is to demonstrate the inherent limits of currently available evidence related to the care of patients with spine related disorders. My discussion will focus on surgical interventions as that is what I have the most experience in; however, the principles apply across all of spine care and indeed all of medicine.

I'd like to start by discussing the care of a young man involved in a motor vehicle accident. He arrives in the emergency department complaining of low back pain. By the time I see him he has had a catheter inserted for urinary retention but strength and sensation are intact in his lower extremities. Computed tomographic (CT) images of his spine are shown in Fig. 1 (Fig. 1). In deciding on a treatment strategy for him, I am able to use what is probably the state-of-the-art classification system for this type of injury, the thoracolumbar injury classification and severity (TLICS) matrix, developed by an expert panel of senior spine surgeons with demonstrated expertise in spine trauma management [1]. According to the TLICS matrix, my patient scores anywhere from a 4 (burst fracture, suspected ligamentous injury, neurologically intact) to an 8 (burst fracture, injured ligamentous complex, cauda equina injury) depending on your interpretation of his facet injuries and urinary retention. Based on this scoring system, I conclude that the best available evidence indicates that he will probably do better if I operate on him than if I treat him nonoperatively. I offer him the option of surgery but he refuses and chooses nonoperative management in a brace. I follow him in the brace and he does very well. His urinary function normalizes once he is mobilized, he requires little or no pain medication, and is home within 48 hours. Six months later, he is completely normal and a follow-up CT shows healing of the fracture and recanalization of the spinal canal (Fig. 2).

Should I be surprised this patient did so well despite a treatment strategy at odds with the best available evidence? In this case, not really. As I mentioned above, the TLICS system is a consensus-based matrix based on an evidence-informed discussion between acknowledged experts. In our hierarchy of evidence-based medicine, this type of evidence falls to the bottom of the heap. Lower quality evidence should be assumed to be imperfect and we should not be surprised if contrary evidence presents itself. That being said, this example does not impugn the usefulness or intrinsic value of the TLICS system—it simply illustrates the fact that imperfect tools are limited in their application even if they are the best tools at hand.

Let us consider another common clinical scenario in which a substantial amount of expert opinion has been thrown about in recent years. The role of fusion following decompression in patients with lumbar stenosis associated with spondylolisthesis has been debated for years. Guidelines published during the past decade dealing with the

https://doi.org/10.1016/j.spinee.2018.11.005

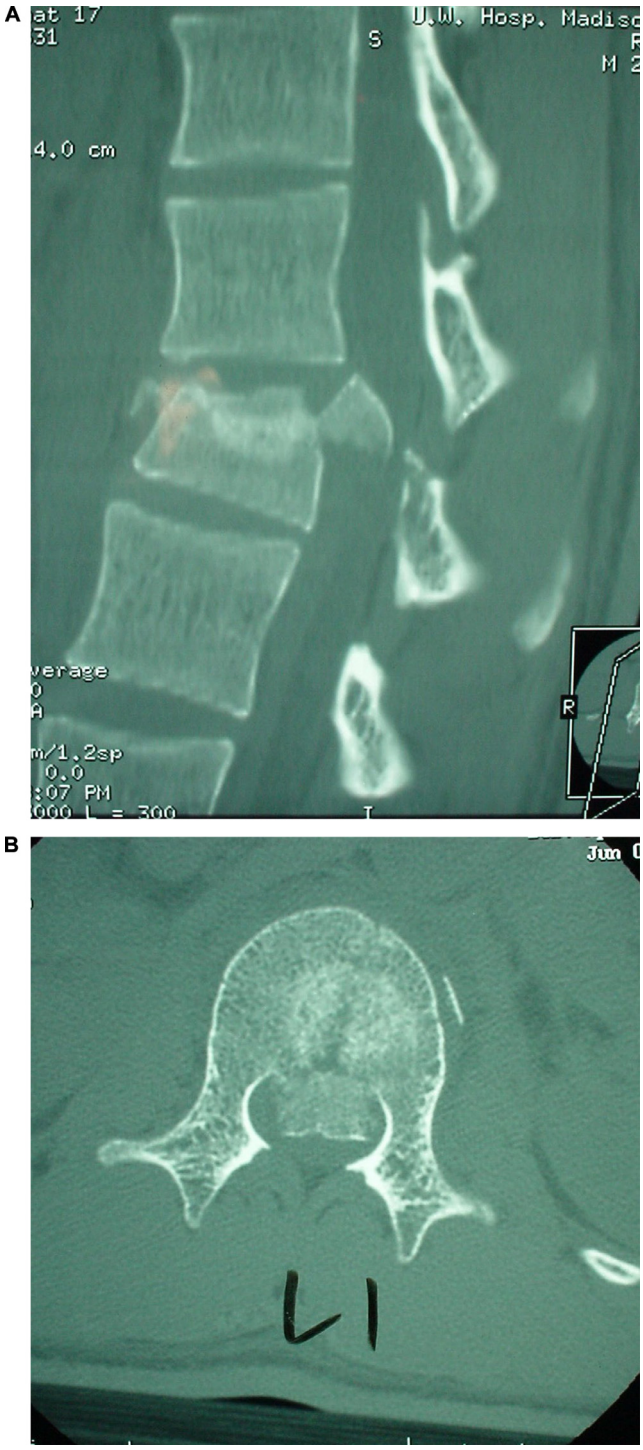

Fig. 1. Images from initial CT in a young man involved in a motor vehicle accident. Sagittal (A) and Axial (B) cuts demonstrate burst morphology, significant repropulsion of bone fragments, kyphotic angulation, and disruption of facets.

issue have generally indicated a role for fusion in this patient population based on acknowledged faulty evidence [2,3]. In 2016, two new randomized trials were published dealing with the exact question of whether fusion was required following decompression in this patient

Fig. 2. Sagittal CT from same patient six months later after management in brace. There has been recanalization of spinal canal and healing of the fracture.

population. Ghogawala et al published the results of a carefully controlled study in which patients with "stable" spondylolisthesis were randomized to decompression alone or decompression followed by fusion [4]. In this study, patient outcomes were improved by the addition of a fusion to the decompression and re-operation rates were lower. A second study examining the role of fusion as an adjunct to decompression in patients with stenosis was published in the same issue of the *New England Journal of Medicine*. This study, authored by Försth and colleagues, examined the role of fusion as an adjunct to decompression in patients with stenosis with or without spondylolisthesis. This study was less well controlled but did consider patients with spondylolisthesis as a separate group. These authors found no improvement in patient outcomes when fusion was added to a decompression in patients with or without spondylolisthesis [5]. How can two randomized controlled studies purportedly looking at the same clinical question come up with exactly opposite results?

There are a few plausible explanations. First, there is math. Depending on the size of the treatment effect, it is more or less possible that a series of identical studies will detect a statistically significant difference between treatment groups. If we set our judgment of "significance" at 1 in 20 (p=0.05), then 1 in 20 studies will be falsely positive, and if we design a study to have a statistical power of 0.8, then 1 in 5 studies will be falsely negative. Second, there is trial design. The two studies cited above are not actually identical. They have different patient populations based on radiographic criteria. Their patients are drawn from different cultures with almost certainly different treatment expectations. They included different sorts of patients (ie, multilevel versus single-level disease) and employed different fusion modalities. Finally, the two studies used different outcomes measures to determine the worth of fusion as an adjunct to decompression for stenosis associated with

spondylolisthesis. It should not be surprising that the studies yielded different results when the differences between the studies are spelled out. At this point in my career, I am convinced that if asked, I could design two randomized studies looking at the same clinical question within the realm of spine interventions which would have exactly opposite results simply by altering factors such as patient selection, details of intervention, outcome measures employed, and timing of outcome assessment. This illustrates the fact that while a randomized trial design can eliminate many biases related to the patient population, it does not do anything to eliminate biases brought to the table by the investigators.

Investigators are human. Because they are human they have biases. Patients who sign up for randomized studies are also human. Because they are human, they also have biases. Editors who choose papers for inclusion in their journals are also human. Because they are human, they are biased. Funding organizations are made up of humans with vested interests in the results of studies they fund. This also introduces bias. The remainder of this paper will be devoted to a description of bias in the spinal literature with a discussion about how best to interpret literature that may be more or less tainted by bias.

Numerous sources of bias can influence the results of a clinical study. One definition of bias in research, ascribed to David Sackett (one of the gurus of evidence-based medicine) is that bias is any factor which systematically influences the interpretation of data away from the truth. Ruth Ellison provides a very nice online synopsis of many of the more common biases seen in trial design (www.uxdesign.cc). In the case of randomized trials, we are really dealing with cognitive biases, both on the part of the investigator as well as the research subject. Common cognitive biases include conformation bias, a tendency to interpret all results as consistent with our original hypothesis even if the results are actually not in line with the hypotheses. The next time you read a journal article, get through the results section, and then are surprised by the conclusions, you will appreciate how common this bias is in our literature. A second common bias is the "bandwagon" effect where investigators reach a compromise in terms of interpretation of data and pretty much stick to the negotiated conclusions in order to avoid conflict within the group. Anchoring bias refers to the tendency to view new information through the lens of previous information. Because of this bias, it is much more difficult to "unprove" something that has previously been proven as all subsequent data are viewed as lesser quality. Think of any one of a series of new technologies introduced by thought leaders with amazing clinical series data. How many negative experiences were required to take the "bloom off the rose" as it were and relegate the technology to the back burner. Clustering illusion is the tendency for humans to look for patterns even when none exist. This is usually operant in the setting of underpowered studies and is avoidable by properly sizing studies for appropriate power.

The most common biases which affect our decision-making capacity relate to trial design (ie, selection bias), expectancy bias, publication bias, and sponsorship bias. As I mentioned previously, choosing the appropriate patient population is critical to the success or failure of a clinical study. One recent example is the Coflex implant which was compared to fusion following decompression in patients with lumbar stenosis (over half of the patients in the study had no spinal deformity) [6]. The spine community has previously determined that fusion is unnecessary in the absence of deformity or instability and that the addition of fusion is associated with higher complications and higher costs without improving patient outcomes [7]. Inclusion of these patients in the study is a clear example of bias being introduced by inappropriate patient selection. Choice of outcomes measures and timing of these measures can also significantly influence the results of a study. If we were, for example, to look at the effect of epidural steroid injections in a population of patients with lumbar stenosis associated with spondylolisthesis, we could easily achieve beneficial results if pain and functional outcomes were measured in the immediate post injection period. These benefits would disappear, however, if we instead measured outcomes at one year following intervention. In contrast, measuring pain outcomes immediately after surgery in the same patient population would likely indicate a harmful effect when longer term outcomes are known to be favorable. When we choose outcomes measures we are also subject to bias. Radiographic fusion is a notorious "false end point" for instrumentation studies due to the tenuous relationship between instrumentation and functional outcomes [8]. Similarly, return to work is an easily measured outcome measure which is of great interest to employers and payers. However, if we are interested in the efficacy of an intervention on back related pain and disability, return to work becomes problematic because many patients with continued pain return to work and many patients with impressive improvements on pain and disability scores never return to work.

Expectancy bias is a two-way phenomenon. It results from the tendency for both investigators and study subjects to influence outcomes through sometimes very subtle social cues. A story used to illustrate the phenomenon comes from turn of the century Austria where Whilhelm von Osten was known for his horse named "Clever Hans." Hans, it was purported, knew math. When presented with even fairly complex arithmetic problems, the horse would tap its hoof the correct number of times to indicate the answer. Subsequent evaluation revealed that Hans really did not know much in the way of math but did know a lot in the way of pleasing von Osten. The horse picked up on subtle cues invisible to most human observers in order to arrive at the "correct" answer. Similarly, when evaluating a patient treated with a new strategy in which an investigator has an interest, the patient may be subtly induced to provide better results than a similar patient treated with a competing

strategy. Blinding the investigator to treatment group can alleviate this effect to some extent but not completely because the study subject also has an expectancy bias. The reason that patients sign up for studies is because they wish access to new, exciting, cool technologies. When they get these new technologies, they are pleased and their outcomes tend to be better than those who were disappointed due to their assignment to the old, out of date, "standard" therapies. Double blinding should largely eliminate expectation bias but double blinding is very difficult to do. Interestingly, however, where double blinding has been performed, interventions have not fared so well [9].

Reporting bias refers to the fact that when experiments go right, they are usually published and when they go wrong, they are not as readily published. This bias affects both authors and editors. Author reporting bias is most pernicious when the author decides not to publish results that are inconsistent with his/her hypothesis. The most egregious examples of this exist in the pharmaceutical literature where it was found that multiple negative studies never saw the light of day despite rigorous design and execution [10]. However, even without direct financial conflicts of interest, academic reputation, personal pride, and fear of ridicule can prevent a researcher from publishing negative results. How many times have we heard the story of a dedicated researcher trying to solve a difficult problem through dozens or hundreds of failed experiments—none of these failures is considered newsworthy but the one experiment that works elevates the researcher's career. This may not be so bad when each experiment is a little different than the last, sometimes it takes time to find the "special sauce" that allows a new development to be realized. However in clinical trials, there really should not be so much variation between trial designs if we are truly trying to replicate the clinical situation.

The situation is made worse by the fact that getting negative studies published is much harder to do than getting positive studies published. Editors of journals depend on metrics such as the impact factor to demonstrate proficiency, attract manuscripts, and improve the relative standing of their journal. Positive studies are cited more frequently, are more exciting, and have a higher profile than negative studies. Multiple authors have described reporting/publication bias in many fields of medicine [11]. There are statistical methods that can be used to document a reporting bias. If one knows the effect size of a particular intervention, one can predict the proportion of studies which would be expected to turn out positive if the results of all studies were reported. When such an exercise is applied to the medical literature, it is clear that there are a disproportionate number of positive studies even when obvious conflicts of interest are accounted for.

Sponsorship bias is extremely important when it comes to medical device, technique, or pharmaceutical trials. As it turns out, when trials are sponsored by manufacturers of drugs or devices, the results of the trials are disproportionately positive even when all other biases are accounted for [12-15]. This phenomenon exists despite the fact that many industry sponsored trials are actually better designed and executed than nonindustry supported trials [16]. The Cochrane collaboration examined this issue and stated that "Sponsorship of drug and device studies by the manufacturing company leads to more favorable results and conclusions than sponsorship by other sources, our analysis suggests the existence of an industry bias that cannot be explained by standard 'risk of bias' assessments [17]."

So then, what are we to do when trying to make decisions regarding the use of new techniques and technology in our patients? We are bombarded with claims of "level one evidence" based on industry sponsored clinical trials, despite the fact that many of these trials have significant patient selection biases, are vulnerable to expectation bias, were performed overseas, and are funded by the manufacturer. Are we to disregard this information? Certainly, information from even a flawed randomized study is better than that derived from case series which were felt (by our community) adequate to introduce interbody cages and trans sacral screws. I personally embrace certain technologies such as image guidance, pedicle screw instrumentation, and minimally invasive surgical techniques which have a relatively weak evidence base in the literature—is this hypocrisy? If we are to demand better information prior to exposing our patients to potentially dangerous and certainly expensive technologies who will provide it? It is already a long and expensive road to get a medical device approved for use in the United States—how high do we want to set the bar and what effect would raising the bar have on innovation? Are we, in fact, asking the impossible?

Many things have been considered impossible. Walt Disney said that "it's kind of fun to do the impossible," and Nelson Mandela stated "it always seems impossible until it is done." That said, I think we need to be realistic with our expectations. We need to continue to challenge our industry partners to improve the quality of the evidence supporting the adoption of new technologies and challenge ourselves to continuously evaluate our practice behaviors and work to improve relevant patient outcomes. As chair of the NASS registry committee, I have refrained thus far from extolling the potential benefits of patient outcomes registries such as the NASS product, but truth be told, a prospective registry employing validated patient-reported outcomes is a perfect mechanism for postmarketing surveillance of drugs and devices, and a perfect quality improvement tool to evaluate the relative efficacy of a broad range of treatment modalities. Taking off the registry chairman hat, I need to acknowledge that our registry products are in their infancy and that it will be years before we'll have meaningful data to influence policy decisions. That said, it doesn't take long at all to influence individual practice decisions. As it turns out, when I reviewed my own experience with fusion for axial back pain, I learned that I am not very good at treating

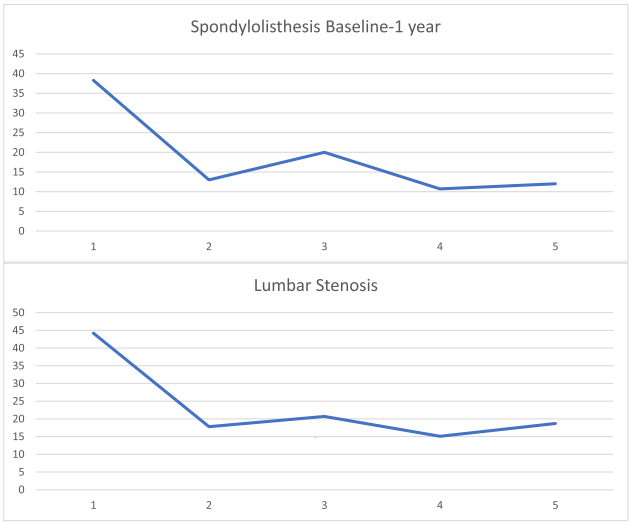

Fig. 3. Graphic representation of Oswestry Disability Index scores for my patients treated for neurogenic claudication with (top) or without (bottom) spondylolisthesis. Time points illustrated are 1. preop; 2. 6 weeks; 3. 3 months; 4. 6 months; and 5. 1 year. This data was obtained using the NASS registry and allows me to follow my patient outcomes and make real time adjustments to my practice patterns in a rational fashion.

axial back pain in the absence of deformity. Conversely, I seem to be really good at treating neurogenic claudication (Fig. 3). I have made a conscious decision to become extremely selective in choosing patients to operate on for axial low back pain. Realistically, if I would not gladly have dinner with you in my home you will not be getting that operation. On the other hand, if you have neurogenic claudication, have had appropriate nonoperative care, and have appropriate imaging findings, I will operate on you pretty much all of the time regardless of most secondary gain issues.

Use of a registry will help in determining the worth of what you are already doing but does not help when making the initial decision to adopt a new technology. What standards should we use when deciding to adopt a new technology? Moses Maimonides was a Jewish philosopher and physician with Egyptian and Spanish roots who wrote a book titled, *Guidelines for the Perplexed* in the 12th century (a rough translation, www.britannica.com). In this book he included an oath for physicians. The principles behind the oath are universal and not limited to any particular culture, region, or religion and have been espoused by Avicenna, Hypocrates, and others. When we are forced to make decisions based on uncertain data, I find it worthwhile to fall back onto these principles for guidance.

> *The eternal providence has appointed me to watch over the life and health of Thy creatures.*

We have been granted an awesome responsibility. We need to continuously ask ourselves if we are awesome enough to live up to these responsibilities. There are way too many examples of spine care practitioners acting badly, way too many examples of inappropriate surgery, way too many of us in the news for tax evasion, Medicare fraud, and other violations of professional conduct. We need to do better.

> *May the love for my art actuate me at all time; may neither avarice nor miserliness, nor thirst for glory or for a great reputation engage my mind; for the enemies of truth and philanthropy could easily deceive me and make me forgetful of my lofty aim of doing good to Thy children.*

Greed does not look good in our profession. I worked with a major television network a few years ago and was presented data indicating that inappropriate surgeries were being done systematically in a very vulnerable patient population. I have also reviewed cases where surgeons were receiving kickbacks for using instrumentation and where miraculous cures were being advertised for obviously sham procedures. This is not acceptable. With regard to new technologies, if the results claimed in an industry study regarding a new intervention seem too good to be true they probably are. Do not be first, do not repeat the same mistakes we as a spine society have made so many times before.

> *May I never see in the patient anything but a fellow creature in pain.*

Our patients are not annuities, they are our parents, our children, and our friends. Treat them as such.

> *Grant me the strength, time and opportunity always to correct what I have acquired, always to extend its domain; for knowledge is immense and the spirit of man can extend indefinitely to enrich itself daily with new requirements. Today he can discover his errors of yesterday and tomorrow he can obtain a new light on what he thinks himself sure of today.*

Honest self-reflection is required to learn what we are actually accomplishing. The systems we deal with are complex and not always intuitive. Burnout is a major issue in our specialty and burn out leads to poor behavior. To avoid burnout we need to participate in the spine community and share our experiences, not just our successes, but more importantly our failures and frustrations. We need to mentor and be mentored. I cannot overstate the importance of mentorship and its beneficial influence on practice. One of the reasons I selected my practice setting was because of the presence of respected, experienced mentors who were not threatened by my presence and who have graciously provided advice over the years. Due to changes largely brought on by the Affordable Care Act, young physicians are more and more frequently employed by hospitals or hospital systems. This results in young physicians being isolated in systems with no intrinsic support, demanding call obligations, unrealistic performance incentives, and no one to talk to on a day-to-day/week-to-week basis. This is a

recipe for disaster that has been realized way too many times. If you are an experienced physician, be a mentor. If you are young physician, get a mentor. If local market forces do not allow this to be in your neighborhood, then reach out.

Oh, God, Thou has appointed me to watch over the life and death of Thy creatures; here am I ready for my vocation and now I turn unto my calling.

It has been an absolute privilege and a pleasure to serve as president of this great organization. I am grateful for the opportunity to advance the mission of NASS and am confident that I leave it in capable hands.

References

[1] Lee JY, Vaccaro AR, Lim MR, et al. Thoracolumbar injury classification and severity score: a new paradigm for the treatment of thoracolumbar spine trauma. J Orthop Sci 2005;10(6):671–5.

[2] Resnick DK, Watters 3rd WC, Sharan A, Mummaneni PV, Dailey AT, Wang JC, Choudhri TF, Eck J, Ghogawala Z, Groff MW, Dhall SS, Kaiser MG. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 9: lumbar fusion for stenosis with spondylolisthesis. J Neurosurg Spine 2014 Jul;21(1):54–61. https://doi.org/10.3171/2014.4.SPINE14274.

[3] Watters 3rd WC, Bono CM, Gilbert TJ, Kreiner DS, Mazanec DJ, Shaffer WO, Baisden J, Easa JE, Fernand R, Ghiselli G, Heggeness MH, Mendel RC, O'Neill C, Reitman CA, Resnick DK, Summers JT, Timmons RB, Toton JF, Society North American Spine. An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spondylolisthesis. Spine J 2009 Jul;9(7):609–14 Epub 2009 May 17. https://doi.org/10.1016/j.spinee.2009.03.016.

[4] Ghogawala Z, Dziura J, Butler WE, Dai F, Terrin N, Magge SN, Coumans JV, Harrington JF, Amin-Hanjani S, Schwartz JS, Sonntag VK, 2nd Barker FG, Benzel EC. Laminectomy plus Fusion versus Laminectomy Alone for Lumbar Spondylolisthesis. N Engl J Med 2016 Apr 14;374(15):1424–34. https://doi.org/10.1056/NEJMoa1508788.

[5] Försth P, Ólafsson G, Carlsson T, Frost A, Borgström F, Fritzell P, Öhagen P, Michaëlsson K, Sandén B. A Randomized, Controlled Trial of Fusion Surgery for Lumbar Spinal Stenosis. N Engl J Med 2016 Apr 14;374(15):1413–23. https://doi.org/10.1056/NEJMoa1513721.

[6] Musacchio MJ, Lauryssen C, Davis RJ, Bae HW, Peloza JH, Guyer RD, Zigler JE, Ohnmeiss DD, Leary S. Evaluation of Decompression and Interlaminar Stabilization Compared with Decompression and Fusion for the Treatment of Lumbar Spinal Stenosis: 5-year Follow-up of a Prospective, Randomized, Controlled Trial. Int J Spine Surg 2016 Jan 26;10:6. eCollection 2016. https://doi.org/10.14444/3006.

[7] Resnick DK, 3rd Watters WC, Mummaneni PV, Dailey AT, Choudhri TF, Eck JC, Sharan A, Groff MW, Wang JC, Ghogawala Z, Dhall SS, Kaiser MG. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 10: lumbar fusion for stenosis without spondylolisthesis. J Neurosurg Spine 2014 Jul;21(1):62–6. https://doi.org/10.3171/2014.4.SPINE14275.

[8] Dhall SS, Choudhri TF, Eck JC, Groff MW, Ghogawala Z, 3rd Watters WC, Dailey AT, Resnick DK, Sharan A, Mummaneni PV, Wang JC, Kaiser MG. Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 5: correlation between radiographic outcome and function. J Neurosurg Spine 2014 Jul;21(1):31–6. https://doi.org/10.3171/2014.4.SPINE14268.

[9] Buchbinder R, Golmohammadi K, Johnston RV, Owen RJ, Homik J, Jones A, Dhillon SS, Kallmes DF, Lambert RG. Percutaneous vertebroplasty for osteoporotic vertebral compression fracture. Cochrane Database Syst Rev 2015 Apr 30(4):CD006349. https://doi.org/10.1002/14651858.CD006349.pub2.

[10] George SL, Buyse M. Data fraud in clinical trials. Clin Investig (Lond) 2015;5(2):161–73.

[11] Francis G, Tanzman J, Matthews WJ. Excess success for psychology articles published in the journal Science. Plos One 2014. https://doi.org/10.1371/journal.pone.0114255.

[12] Leopold SS, Warme WJ, Fritz Braunlich E, Shott S. Association between funding source and study outcome in orthopaedic research. Clin Orthop Relat Res 2003;415:293–301.

[13] Flacco ME, Manzoli L, Boccia S, Capasso L, Aleksovska K, Rosso A, Scaioli G, De Vito C, Siliquini R, Villari P, Ioannidis JP. Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. J Clin Epidemiol 2015;68(7):811–20.

[14] Naci H, Dias S, Ades AE. Industry sponsorship bias in research findings: a network meta-analysis of LDL cholesterol reduction in randomised trials of statins. BMJ 2014;349:g5741.

[15] Pang WK, Yeter KC, Torralba KD, Spencer HJ, Khan NA. Financial conflicts of interest and their association with outcome and quality of fibromyalgia drug therapy randomized controlled trials. Int J Rheum Dis 2015;27(10):12607.

[16] Wareham. Sponsorship bias and quality of randomised trials in veterinary medicine. BMC Veterinary Research 2017;13:234. https://doi.org/10.1186/s12917-017-1146-9.

[17] Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev 2012;12:MR000033.