Clin Oral Invest (2013) 17:1789–1790
DOI 10.1007/s00784-013-1020-5

LETTER TO THE EDITOR

Trial quality checklists: on the need to multiply (not add) scores

Kaitlin E. Palys · Vance W. Berger · Sunny Alperson

Received: 26 July 2012 / Accepted: 10 June 2013 / Published online: 22 June 2013

Editor:

We applaud Craane, Dijkstra, Stappaerts, and Laat [1] for their efforts in comparing and contrasting the Delphi, Jadad, Risk of Bias, and Megens and Harris lists for evaluating trial quality, and we admire their hesitation in selecting one of these as optimal. However, limited areas of disagreement do exist. First, although the authors never claimed explicitly that these four lists are the only ones available for assessing trial quality, they do seem to imply this false dichotomy (or quatrotomy, as the case may be) by limiting consideration to them and asking which one is best. In reality, there are other lists, including the more comprehensive Chalmers list [2].

Let us consider as an analogy a walking bridge with 100 wooden planks, any one of which may be defective. The chain truly is as strong as its weakest link, and if we decide to trust this bridge, and walk over it, and if even one plank is in fact defective, then we end up in the river below. One safety inspector wants to check every third plank; another, every fourth; another, every fifth; and yet another wants to check every seventh. As we noted in our first point, these are not the only options, but let us bear this analogy in mind as we continue.

We note the unfortunate wording in stating that "The Delphi list scored significantly lower than the other lists." The casual reader comes away with the impression that the Delphi list is the worst among the lists considered, whereas only the reader who can keep his or her eye on the ball recognizes that the list does its job when it uncovers problems with the trials. Low scores, while damning for the trials evaluated, actually indicate that the list did its job and uncovered problems. We would not criticize a safety inspector for finding safety violations, or faulty planks, that other inspectors missed. This is a good thing: inspect more planks, not fewer.

So it is clear [3, 4] that a list is ideal not when it is efficient or quick to use, but rather when it is comprehensive. Hence, the very question of which one is best is a red herring. Each of us has benefited greatly over the years from sound advice from a variety of mentors. Though we may reflect, every now and again, on which mentor was best, we also recognize this question to be of academic interest only, and we never ask ourselves if we are better off following this advice or that advice. When presented with good advice from a variety of sources, we combine these into something new, something better, something more comprehensive. Do we inspect every third plank, or every fourth plank? What about inspecting both sets of planks? As a Venn diagram would illustrate, no list can be universally best unless it contains all the others. In general, the best list one can derive to from a given set of lists would be the union of all of these lists. That is, we would create a new list that would include anything that appears on any of the input lists. This option, not considered by Craane et al., would yield a criteria list that is more comprehensive than any individual list. The union misses fewer planks. This is a good thing.

If none of the lists contains an "other" element [4], then the union will not have one either; hence, even the union cannot be comprehensive. This is not a direct criticism of Craane et al. but rather of the quality lists themselves. Although these criteria may address crucial factors regarding the worth of a trial, unforeseen circumstances could arise that render a trial

K. E. Palys (✉)
Virginia Polytechnic Institute and State University,
Blacksburg, VA, USA
e-mail: palysk@vt.edu

V. W. Berger
Biometry Research Group, National Cancer Institute, and UMBC,
Executive Plaza North, Suite 3131 6130 Executive Boulevard,
MSC 7354, Bethesda, MD 20892-7354, USA
e-mail: vb78c@nih.gov

S. Alperson
Montgomery College, Takoma Park, MD, USA
e-mail: sunny.alperson@montgomerycollege.edu

fatally flawed, but if the list lacks an indicator for that given situation, then the score for the trial will still be (artificially) high. Sadly, it is precisely *because* certain lists, most notably the Jadad score, consider so few dimensions of trial quality that they are selected in practice [3, 5].

Any plank can be defective. Arguing over whether to sample more or less is a disservice and, most certainly, is not progress; inspect all of them. Beyond that, the connections between them may also be defective, as can the support structures; add in an "other" element in case all is not what it seems. Go back to the Chalmers list [2] as the gold standard, do not cut corners by considering these more minimalist lists, and when presented with a set of lists, go with the union. Check every element of each list, plus look around for other problems not anticipated by these lists. This is how science should be done. But this still leaves one question unanswered. Once we have our comprehensive union, with its large number of elements, how do we derive an overall quality score? Craane et al. followed the conventional wisdom of summing the individual scores. So when we inspect every plank and find 97 solid, and only three rotted out, we score 97 %; that is not too bad. When a victim of a fatal heart attack happens to have healthy lungs, kidneys, and other organs, we might still use this additive logic to compute a high overall health score. The reality is that the failure of only one organ can kill you, and one faulty plank can also kill you. One fatal flaw in a trial can kill a trial, and there is no compensation in getting everything else right [5]. Each element scores a 0 or a 1, and then these scores are multiplied, not added, so that even a single 0 reflects the fact that the trial is fatally flawed [5].

Our views on the evaluation of trial quality reflect our perspective that this exercise should be undertaken in a scientific manner for the good of the patients who will rely on the future medical decisions informed, to a greater or lesser degree, by the trials whose quality is being evaluated. If others value convenience to the evaluators over the safety of future patients, and may also have conflicts of interest that cause them to want to assign perfect scores to the quality of the trials they evaluate, then it would not be at all surprising that they would be more interested in how quickly they can rubber-stamp a perfect score, and hence would want to use the Jadad score. The question becomes what is better for society. The answer should be clear.

References

1. Craane B, Dijkstra PU, Stappaerts K, De Laat A (2012) Methodological quality of a systematic review on physical therapy for temporomandibular disorders: influence of hand search and quality scales. Clin Oral Investig 16(1):295–303
2. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D (1981) A method for assessing the quality of a randomized control trial. Control Clin Trials 2(1):31–49
3. Palys K, Berger VW (2012) A note on the Jadad score as an efficient tool for measuring trial quality, J Gastrointest Surg 17(6):1170-1. doi:10.1007/s11605-012-2106-0.
4. Berger VW, Alperson SY (2009) A general framework for the evaluation of clinical trial quality. Rev Recent Clin Trials 4(2):79–88
5. Berger VW (2006) Is the Jadad score the proper evaluation of trials? J Rheumatol 33(8):1710