Journal of Scientific Exploration, Vol. 26, No. 3, pp. 635–646, 2012 0892-3310/12

ESSAY

The Review Reviewed: Stop Publication Bias

J. Alexander de Ru

Department of Otorhinolaryngology and Head & Neck Surgery
Central Military Hospital, Utrecht, The Netherlands
j.a.deru@umcutrecht.nl

John C. M. J. de Groot

Department of Otorhinolaryngology and Head & Neck Surgery,
Leiden University Medical Center, The Netherlands

Jan-Willem M. Elshof

Department of Surgery,
St. Elizabeth Hospital, Tilburg, The Netherlands

Submitted 12/27/2011, Accepted 3/1/2012

Abstract—This manuscript describes our past experiences with reviewers and the review procedures that are currently used in the medical sciences. We conclude that reviewers all too often are biased, whereas scientific discussion should be based on substantive comments and without prejudice. In our opinion, subjective arguments for rejection of manuscripts constitute a serious threat to evidence-based medicine. Since peer review should aim to facilitate the introduction into medicine of improved ways of curing, relieving, and comforting patients, a more objective review system with greater scope for the publication of divergent opinions is clearly needed to ensure that a literature search does not merely produce a plethora of articles with mainstream opinions. Our recommendations for a peer review system are: (1) No more anonymous reviewers; (2) The reviewer must concentrate initially on two questions: (a) was a real problem formulated in this manuscript? and (b) is the conclusion—if proven—relevant for practical situations?

Keywords: publication bias—evidence-based medicine—review

Introduction

Having a good, well-structured idea is one thing; getting it published is something else entirely. In science, however, the first should—more or less automatically—lead to the second. It has long been recognized that scientific revolutions meet stiff resistance. The battle that the 2011 Nobel Laureate for Chemistry, Daniel Shechtman, had to fight for years against established

science is only a recent albeit shocking example. This phenomenon has already been described long ago by Thomas Kuhn (1970), and Juan Campanario (Campanario & Martin 2004, Companario 2009) has published many examples of the resistance encountered by prominent physicists in gaining acceptance for their theories. If the problems encountered in the exact sciences are so great, what then are the prospects for medical science?

The Laureates of the 2005 Nobel Prize for Medicine had to infect themselves with the bacterium *Helicobacter pylori* in order to prove its role in gastritis and peptic ulcer disease. An idea borrowed from John Hunter, who in 1767 inoculated himself through the urethra with pus from a gonorrhoea patient. Why have we learned so little from the past? It seems that reviewers and their reasons for opposition will never change. And not just when it relates to paradigmatic changes, but even more so when it concerns more modest scientific findings. This means that other measures are required to ensure that medical science becomes more open to new and divergent ideas.

Another reason for tackling the problem of incomprehension and even opposition to such ideas is that it leads to publication bias. Publication bias is defined as "the tendency on the part of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings" (Song et al. 2010). The first step toward the prevention of publication bias is to make the public aware of the sometimes detrimental effects of publication bias and the need for the results of all studies to be made accessible (Song et al. 2010).

We agree, of course, that evidence-based medicine (EBM), if correctly applied, is a highly logical and systematic approach to clinical practice. However, it is conditional on two constraints: (1) all evidence needs to be in the literature; and (2) all levels of evidence must be evaluated. The second of these often tends to be forgotten by many assessors for guidelines and systematic reviews. The main problem, however, and one which does not get much attention, is that not all evidence ends up in the literature. This is what we want to discuss here, because the validity of research synthesis based on published literature will be threatened if published studies comprise a biased selection of all the studies that have been conducted.

For what is the use of the usual literature studies in EBM if the literature on which they are based is a collection of mainstream works which seems to have come about more by "mutual admiration" rather than by peer review? Evidently, people prefer to see many investigators taking little steps along the beaten track rather than fewer investigators with divergent and sometimes pioneering ideas. That is a pity, since, as Proetz (1964) once wrote:

> Perhaps at first glance it may seem immaterial whether science is advanced by a handful of geniuses in leaps and bounds, or by a million average citizens in creeps and budges, but there is a waste of both money and manpower, a spotty distribution of knowledge, and the necessity of having to weed out the errors and start over.

What is the point of weighing up the pros and cons if you can only find the pros because the cons have not been published? As Evans (1995) pointed out:

> It is to use evidence in the manner of the fabled drunkard who searched under the street lamp for his door key because that is where the light was, even though he had dropped the key somewhere else.

We will show, also from our own experience, that the review process is all too often a subjective activity and that rejections are frequently made on improper and hardly scientific grounds. We do not get frustrated by well-argued rejections, and we are also well aware that journals get sent far more copy than they could ever publish. We are aware that even the examples given below are debatable. All we ask is that a scientific discussion be based on real substance and without prejudice.

We would argue that an appraisal by a reviewer requires a quite different strategy from the assessment of an already published article for a guideline, a critically appraised topic, or a systematic review. We consider that the review process should aim to publish a number of different insights in order to avoid ending up with *evidence-biased medicine*.

Manuscripts and Reviewers

Title and References

The first place where things can go wrong in a manuscript appraisal is with the title and the references. The requirements for a title differ according to the journal. Should it be gripping or businesslike? How long should it be? Plenty of advice is available on this matter (Fraser 2008, Hall 2008).

Criticism of the title can even contribute to a rejection. We believe that it is up to the Editor to ask for a change of title if he thinks it necessary. It is not the task of the reviewer to make more than a minor comment on this.

Many journals ask authors to limit the number of references. Only cite relevant (and possibly contradictory) articles in order to place your story in its context, not to demonstrate how much you have read. Back in 1964, Proetz already wrote that there was far too much redundant name dropping and that it would be refreshing to read something in medicine without being

constantly interrupted by what Kussmaul and Rokitansky wrote about it in 1878.

Reviewing the contents on the basis of the title and references, as sometimes happens, is—in our opinion—like judging a wine on the basis of the label on the bottle.

Length and Form

How long or how short should a manuscript be? When have you given too little background information and when too much? Differing opinions on this matter should not form grounds for rejection. Of course, an article should be written in a short and concise style, but we believe that it eventually should be the task of the Editorial staff of a journal to judge this and, if necessary, make improvements.

Moreover, the length of a manuscript can give an impression of the underlying thought. A two-page commentary is not a trial and should not be judged as such. An article can sometimes be too long for the regular version of a journal, but this should not form a problem for a supplement.

Another factor is the form, the "personal" writing style. It is, of course, advisable for authors, especially if they are not native speakers, to have their manuscripts textually and linguistically checked by others. But the reviewer could also make suggestions for improvement, and journals could and should make a greater contribution. After all, some authors write better than others. But this is quite a separate issue from the significance of the message.

Contents, Type of Investigation, and Level of Evidence

Should all articles be subjected to the same appraisal? We think not. A case report or an opinion article is not a systematic review and should not be judged as if it were. We do not believe that the recommendations, as described in the CONSORT statement, are relevant here, and reviewers should realize that the overzealous application of rules, which were drawn up in connection with recommendations for the proper performance of randomized controlled trials (RCTs), to completely different sorts of investigation is quite absurd. Small-scale retrospective studies are recorded after the patient has left, and the procedures followed can no longer be changed. Needless to say, by then you no longer have all the relevant information available. Retrospective studies, opinion articles, and narrative reviews also have a practical role, especially for theoretical orientation. [There is no need here to mention the practical shortcomings of RCTs, since enough has been written on this subject in the past (Kaptchuk 2001, Rothwell 2005).]

In the final determination of the level of evidence needed for a guideline or systematic review, however, proper account must be taken of the methodology and the type of investigation.

Expert opinion is difficult to substantiate with systematic references, but is not without significance. Furthermore, it is important to note the difference between an expert who just makes an unsubstantiated statement and an expert opinion based on literature.

An expert who gives an opinion based on studies at a particular level of evidence can, in fact, be considered as giving an expert opinion at *that same level*. (Usually in EBM, systematic reviews and large trials are considered level 1, cohort studies level 2, case series level 3, and expert opinion as level 4. But an expert giving his opinion based on results from a level 1 trial or review might perhaps better be viewed as level 1 instead of level 4.) It is the ultimate combination of science and specialist know-how, in other words of EBM:

> The practice of EBM means integrating individual clinical expertise with the best available external clinical evidence from systematic review. By individual clinical expertise, we mean the proficiency and judgement that individual clinicians acquire through clinical experience and clinical practice. (Sackett et al. 1996)

So why not assign a due level of evidence to this? If expert opinion is totally worthless, you might—for certain medical problems for which there is no other useful evidence available—just as well ask the gardener as ask a doctor with many years' experience.

Knowledge and Practical Use versus Methodology

How do you get a divergent viewpoint, based on years of experience, published? Nowadays, many medical students are better able to judge the methodological quality of an article than its clinical relevance. Reviewers who are excellent in statistics, epidemiology, and conducting systematic reviews and guidelines have sometimes spent too little time gaining practical skills and experience. Doctors are not statisticians, and it is hardly possible to excel in both fields. Insufficient knowledge of pathology, anatomy, pharmacology, and pathophysiology, and also a lack of clinical experience, can lead someone to reject interesting studies on the grounds that they employ "poor methodology." This quite apart from the fact that—according to statisticians—the statistics in many published articles often leave a lot to be desired and the statistical knowledge of many doctors is evidently inadequate (Ioannidis 2005).

This major problem also affects many systematic reviews and guidelines. Methodologically well-designed studies are presented, while the "flaws" in the definitions used in the Introduction and the Discussion cannot be recognized by non-experts, who are primarily the ones who carry out the literature search. However, just like Greenhalgh (2010), we believe that when interpreting a study, readers need to know how it relates to existing knowledge.

Many authors and reviewers interpret findings narrowly, failing either to identify previous studies or to place their findings in the context of previous studies.

Greenhalgh (2010) wrote that she was concerned that courses in EBM often concentrate too much on critical appraisal and apply insufficient critical evaluation to the other steps: "Yet if you have asked the wrong question or sought the wrong answers (from the wrong sources), you might as well not have read any papers at all." That is like asking a restaurant dishwasher who has read a few cookbooks to prepare a meal. Although the meal might contain all the right ingredients, it is questionable whether it would taste very good.

We should like to take as an example the case in which it was recognized that a prospective study of the effects of Botox in the *corrugator supercilii* muscle was not a double-blind randomized trial, but the fact that Botox has an excellent and clearly visible effect on the face and therefore cannot be blinded was not recognized. In short, you only need to write that an effective randomization was applied and that independent-effect appraisers were employed, in order to be free from any criticism of the medical content.

In addition, it is worth quoting the long-standing Latin aphorism *ubi pus, ibi evacua* ("where there is pus, evacuate it"). There has not been much research into this matter during the last century, probably because this is something that every doctor is expected to know. On the other hand, plenty of research has been done into treatment by means of antibiotics. We now find that we often have to persuade colleagues to operate for example in cases of mastoiditis, even when it has caused meningitis. This is because—according to these colleagues—there is no evidence that an operation is useful, whereas there is evidence that antibiotics can often, eventually, solve the problem. Of course, an antibiotic therapy can have an added value, but we must not misplace the burden of proof. Although in cases with not very ill patients we have no objection to starting with a good second-best treatment, it is questionable whether antibiotics as a single therapy are as effective as the evacuation of the pus combined with antibiotics. The findings of a study into this subject should not only be based on mortality or ultimate cure, but should also take account of complications, morbidity,

length of hospitalization, the duration of antibiotic administration with the concomitant complications, and on the other hand the complications that could arise from an operation. Until genuine comparative evidence to the contrary is produced, we consider that expert opinion for evacuation—combined with antibiotics—will continue to be the gold standard. We would do better to base our decisions on the collective experience of thousands of clinicians treating millions of patients. (In other words: 50 million years of evolution and 50 years of research have demonstrated that mothers' milk is healthy for babies.) Of course, if the area of operation would involve excessive risk of operative complications due to anatomical circumstances, then the preferred treatment would shift to antibiotics. In that case, the treatment of second choice might then be the best option.

The same sort of problem is faced by Ridge (2010), to judge from his article "We show pictures, they show curves." The fact that radiotherapy and chemotherapy, possibly in combination, can constitute good modes of treatment in certain cases does not mean that surgical intervention is not equally effective in the case of tumors of the head and neck. Years of surgical experience by the real experts give results that cannot be expressed in simple graphs.

Another factor is that reviewers are sometimes unable to conduct certain operations themselves. There is no shame in this, since specialist operations, as the term itself implies, cannot be done by everyone. Unfortunately, these reviewers have the habit of attributing other people's good results to mere luck or to the placebo effect. Or even worse, the reviewer asks himself: "Surely the authors do not really intend to . . .?"

Such reviewers consider an endoscopic browlift, a microvascular decompression, or the severing (neurectomy) of the vestibular nerve as excessive techniques and the authors as "trigger happy." But just because you do not yourself perform a particular operation, you cannot leave patients to suffer lifelong pain, facial spasm, or dizziness while relatively routine operations exist, which—in experienced hands—could solve the problem. The decision as to the most appropriate treatment should be taken in the context of clinical practice and not by incompetent reviewers. After all, it is not the reviewer's task to judge what is onerous for the patient. This is a matter for individual doctors to discuss with their patients.

Another source of annoyance in some cases is reviewers' clear lack of general knowledge, although this is usually clothed in the suggestion that the concepts will be unfamiliar to the journal's readers.

Bradford Hill criteria? "Never heard of it." And these people have not even taken the trouble to find out that these are criteria drawn up by one of the foremost epidemiologists of the last century. How can you call yourself

a reviewer if you possess so little knowledge? How is it possible for a doctor to be unable to make any sort of link between smoking and cancer? And what sort of reviewer are you if you will not take the trouble to follow up references?

There is no shame in turning down a request for a review if you do not consider yourself sufficiently expert for the task. Unfortunately, this is not generally recognized.

And then there are the reviewers without an understanding of Bayesian statistics. No shame in itself, but no reason to reject an article; as if Bayesian statistics is something from another planet. They should ask a statistician to give an opinion on the subject if they do not understand it themselves. And if the readers and reviewers are really so ignorant, perhaps they should take things on trust.

Rules for Reviews

Distinguishing genuine research from poor-quality endeavors of well-meaning amateurs is the primary task of reviewers in EBM (Greenhalgh 2010). Naturally, the goal is to aim at high-quality research, but this also—and above all—implies high-quality ideas or theories. Good research should be judged not only on the quality of the methodology, but also on the merits of the research goals and their relevance.

And do these well-meaning amateurs really have such bad ideas, or do they lack the extensive research facilities (including statistical and epidemiological advice) available to academic institutes? For it is a well-known problem that conducting research is becoming more and more difficult for a peripheral clinician (Warlow 2005).

Some time ago, we received a request to review a manuscript accompanied by the specific request to refrain from making comments that we would not like to receive ourselves. An excellent piece of advice that all journals would do well to adopt, since it is all too apparent that not all reviewers do this of their own volition. They could have read in *How to Write a Paper*: “Be kind—it’s a privilege, be helpful, and above all be fair and honest” (Hall 2008).

A reviewer recently insulted us with text such as: “undergraduate university education has passed the authors by,” “amateurish step backward,” “hidden agenda of discouraging some Dutch ORLs [otorhinolaryngologists] from doing excessive numbers of surgeries for personal gain,” and “would harm the high and growing reputation of the Journal.” If this reviewer had furnished some factual and substantiated comments, he might have demonstrated what worthless authors we were and would not have needed to

couch his opinions in such terms. In view of the fact that his comments had absolutely no bearing on our line of reasoning, references, or conclusion, they must have been purely a personal attack by a frustrated, arrogant colleague. An additional factor was that by accepting these comments, despite our complaint, the Editor was indirectly insulting the other reviewer, who had given it a good review.

We believe that a reviewer should no longer be allowed to give a personal opinion unless the anonymity of the review process is removed. Anything which you only dare say behind the veil of anonymity should not be said at all. Subjective, unsubstantiated opinions of reviewers are really not relevant. For example, we once had a reviewer who thought that he would have chosen different parameters in an appraisal system for the degree of facial dysfunction, but if this is not backed up by sound arguments, it is no more than an irrelevant remark. If a reviewer wishes to express a personal opinion, he should write his own article. At the most, he should confine himself to asking why something specific has not been included in the system and suggesting that we comment on this in the Discussion section.

Another point to be observed by a reviewer and which, in our opinion, should be included in the request for review is: no hair-splitting about relatively irrelevant matters. Try, rather, to imagine why a particular concept has been chosen. If someone mentions a follow-up period of 3–30 months, it is a non-issue to point out that it should be at least four months, all the more so if there is no evidence from the literature that four months is the gold standard and this applies to many patients in that situation in any case. One could perhaps take the results slightly less seriously or, preferably, ask the authors to explain why they chose for the one rather than the other. A rejection on the basis of this one month is quite illogical.

If authors have written a retrospective study, it is pointless to ask for prospective results. If relevant, the authors could be asked to comment on this in the Discussion. The possibility of carrying out a prospective study in the future will then remain open to any doubting reviewer—or ultimately the reader—if the results are called into doubt.

Back in 1964, Proetz stated that he considered the most important part of an article, i.e. what you should read first, to be the conclusion. If the conclusion is not worthwhile, then there is no point in reading the rest of the manuscript. Conversely, the more meaningful, innovative, or disparate the conclusion, the more reason there should be to publish. After reading the conclusion, the reviewers should first evaluate the medical content with respect to accuracy and relevance, and only then is the methodology worth appraising—if necessary in collaboration with statisticians, epidemiologists,

and/or other outsiders. We consider the first priority should be to present all opinions, from all levels of evidence. Only after this has been done—in this particular order—is it possible to rank the pros and cons, and only then will the public be well-informed about all the different opinions. This would also make it possible, in certain well-justified cases, to deviate from protocols drawn up on the basis of EBM.

We would prefer not to use formulas, such as the Fail-Safe Number (FSN), to correct for the possibility of not all the evidence being available (Orwin 1983). (And to think that Orwin used this FSN to correct for the fact that studies that do not demonstrate any effect might be more difficult to get published. The fact that controversial opinions are actively boycotted is particularly difficult to express in formulas.)

There is also a generally accepted opinion that the vast majority of research is published in low-impact journals, where peer review is undoubtedly less thorough (Greenhalgh 2010). In the field of otorhinolaryngology (ORL), we would make the point that we have read better review comments on articles we have published in *B-ENT*, a journal positioned somewhat lower in the journal ranking, than those given by many higher-ranked journals within the field. We would almost turn the argument around. The major journals ask big names to conduct their reviews, but these reviewers, with their often big egos, sometimes have difficulty in distinguishing between an objective review and their own personal opinion.

We consider that Editors would do well to take criticism of reviewers seriously and to be more open to asking for second opinions if prompted by justified questions from authors. Greenhalgh (2010) described decision-making by GOBSAT (good old boys sat around a table) for the purpose of arriving at a guideline. This process seems to be even worse when applied to the publishing of a journal. If complaints are made about a reviewer, it is not unusual to be fobbed off with remarks such as "the reviewer is one of the best minds in the business" or "the reviewer is this country's leading expert." As if this is of relevance when someone rejects a manuscript purely on account of prejudice and without sound arguments. Unfortunately, it seems that an Editor seldom overrules his Associate Editor. Whereas it is quite normal in legal proceedings for a higher court to overturn a decision, medical science continues to cling to an unshakable belief in a colleague's infallibility. Even worse is when a request for a second opinion is rejected by the same Associate Editor who had examined the review in the first place, so that the case never gets to be examined by another person, possibly higher in the hierarchy. (In a legal setting, such a thing would be virtually inconceivable in the civilized world.)

Unfortunately, it also happens all too often that articles from leading

journals—chosen because of their titles and the impact factor of the journal—are cited inappropriately, i.e. where they are not specifically applicable, or are misused in practice. Editors should promote self-correction in science and participate in efforts to improve the practice of scientific investigation by publishing corrections, retractions, and letters critical of articles published in their own journal (Altman 2002). That applies, in our opinion, not only to errors in the article, but also—and in particular—if an article is repeatedly misused.

One of the most important messages in this manuscript is to point out to Editors and particularly reviewers that progress in knowledge is best achieved by debate. This is initiated by the publishing of divergent ideas and theories. Once such a divergent theory has been published, it can be objectively evaluated by colleagues in the same specialist field by means of a process of verification or falsification. If divergent ideas are (deliberately) written off, peers or confrères will never even have the opportunity to take cognizance of them.

In short, we believe that it should be made less problematical to get solidly substantiated, divergent opinions published. Our recommendations for a peer review system are:

(1) No more anonymous reviewers
(2) The reviewer must concentrate initially on two questions:
(a) was a real problem formulated in this manuscript?
(b) is the conclusion—if proven—relevant for practical situations?

If the answers are in the affirmative, the study should stand a good chance of being published. At this stage, the reviewer can, of course, ask critical questions about the methodology and request elucidation on ambiguities.

Peer review must aim to facilitate the introduction into medicine of improved ways of curing, relieving, and comforting patients. The fulfillment of this aim requires both quality control and the encouragement of innovation. If an appropriate balance between the two is lost, the peer review will fail to fulfill its purpose (Horrobin 1990). Or, as MacAuley (cited in Hall 2008) wrote: "What matters is originality, importance, and validity." And, we would like to add, in that specific order!

Conclusion

A more objective review system with greater scope for the publication of divergent opinions is needed to ensure that a search through evidence-based medicine does not merely produce an accumulation of articles with a

mainstream opinion and with a mainstream conclusion. The present, overly subjective system leads to publication bias.

Acknowledgment

We would like to thank Esther Caspers for her helpful comments on this paper.

References

Altman, D. G. (2002). Poor quality medical research. What can journals do? *Journal of the American Medical Association*, *287*, 2765–2767.

Campanario, J. M. (2009). Rejecting and resisting Nobel class discoveries: Accounts by Nobel laureates. *Scientometrics, 81*, 549–565.

Campanario, J. M., & Martin, B. (2004). Challenging dominant physics paradigms. *Journal of Scientific Exploration, 18*, 421–438.

Evans, J. G. (1995). Evidence-based and evidence-biased medicine? *Age and Ageing, 24*, 461–563.

Fraser, J. (2008). *How to Publish in Biomedicine*. Oxford: Radcliff Publishing.

Greenhalgh, T. (2010). *How to Read a Paper: The Basics of Evidence-Based Medicine*. Chichester: Wiley-Blackwell.

Hall, G. M. (2008). *How to Write a Paper*. London: Blackwell Publishing BMJ Books.

Horrobin, D. F. (1990). The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association*, *263*, 1438–1441.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Kaptchuk, T. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology*, *54*, 541–549.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (second edition). Chicago: University of Chicago Press.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159.

Proetz, A. W. (1964). How to smother an idea. *Annals of Otology, Rhinology, and Laryngology*, *73*, 364–369.

Ridge, J. A. (2010). We show pictures, they show curves. *Archives of Otolaryngology and Head & Neck Surgery*, *136*, 1170–1175.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this apply?" *Lancet*, *365*, 82–93.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence-based medicine: What it is, and what it isn't. *British Medical Journal*, *312*, 71–72.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, *14*, 1–217.

Warlow, C. (2005). Over-regulation of clinical research: A threat to public health. *Clinical Medicine*, *5*, 33–38.