Clinical evidence

Quality of evidence published in clinical journals: signs of change?

R G Newcombe, J Stebbing

This commentary provides a consensus view by two reviewers on a paper appearing in this issue that investigates the change in the quality of evidence published in clinical journals over a period of 25 years.

John Mayberry sought the opinion of Robert Newcombe as the statistical referee on the accompanying paper¹ after it had been revised twice. At this stage, differences of opinion between the authors and the original referee, Justin Stebbing, remained unresolved. Robert suggested that the paper should be published as it stands, accompanied by the reviewer's comments and his appraisal of the situation. The referees acknowledge that (1) all papers have deficiencies and can be improved, (2) it is meritorious to publish original work that advances a field, and (3) such work is often useful even if it generates more questions than answers. The comments that follow are a consensus view arrived at by these two reviewers.

The paper did not seem a particularly strong contender for journal space. Straightaway, the reference to a 30 year perspective, meaning 1983–2003, did little to inspire confidence. Here, it is important to bear in mind that the whole scientific publication system is based on trust—the reader cannot possibly verify everything that is presented in an article, nor indeed can a referee. The refereeing process is largely a negative one, looking for obvious deficiencies that need correcting, rather than providing positive reassurance that what has been done is appropriate. Moreover in the domain of quality control, retrospective is no adequate substitute for prospective. Furthermore, the paper seems to fail the “So what?” test. Nevertheless, it provides a useful springboard to reflect on the research and publication system in more general terms.

The main areas of dispute seem to be the range of years studied and the statistical methods used. To deal with the latter first, discussion has centred on issues such as choice of statistical test, multiplicity of comparisons, and whether a multivariate analysis adjusting for confounding factors is called for. But there is a more basic issue. In any

piece of research, the analyses presented as primary are to correspond directly to the aims and objectives of the study. When stated as a principle, this is self evident. But no aims or objectives are stated in the article—as such there is no “primary end point” (or secondary one for that matter) and a large number of questions are being asked, thus some by virtue of chance will pass a significance test. If the researchers' aims are to record how the characteristics of published articles have changed over the period, then a simple analysis at descriptive level is appropriate. If the aim is to “go deeper” and make interpretations that take into account other factors that have changed, this is when it might be appropriate to use multivariate methods to make adjustment. While many would request this, it would be all too easy to show whatever one wanted to show, simply by judicious selection of which confounding factors to adjust for.

One measure of alleged “quality” considered in the article is the bibliometric impact factor, albeit in relation to the aggregate of articles published, not directly to individual articles. Impact factors of journals are published annually (not all journals, but a selected elite group). The impact factor is an extremely crude index of impact on research, and continues to be used only because it is virtually impossible to devise any better index; moreover it does not necessarily equate with impact on clinical practice. It has played an important part in several rounds of the Research Assessment Exercise that has acted to the undoubted detriment of academia in the UK² and forms a large part of author's considerations when submitting a manuscript.

In terms of adding more datapoints, we would expect results for 2004 to be very similar to those for 2003, the difference being largely a matter of random variation. Including information for 1988 and 1998 would provide a fuller picture, although whether it

would be worth the additional effort seems questionable. Obviously what would add the most interest would be to include data for 2013, but we would have to wait a long time for this. Arguably, the data for 1983 would then seem of limited relevance, just as the authors felt about 1973.

Whichever years are selected, the fact remains that a historical comparison is being made. Many things have changed beyond recognition during the period of study (although this should not be regarded as a prompt to perform the multivariate analyses mentioned above). For instance, in 1983, the HIV epidemic and targeted oncology (neither infectious diseases nor cancer are included in their analysis) and the systematic review industry (the absence of meta-analyses from the 1983 series is hardly surprising) were in their infancy, and the internet as we now know it did not exist. The null hypothesis for some of the statistical tests in this article—that proportions of different kinds of studies do not change over the study period—is a non-starter. It also came as no surprise to us that these proportions differ astronomically significantly between specialties—we would have been most surprised had they turned out to be equal.

Counting articles, as here, may be of no greater validity than the way in which the impact factor is constructed, nor the simplistic use of qualitative analysis software to report counts of appearances of particular words in source text. Figures cannot exclude multiple publication of data from the same study unless the investigators have specifically and scrupulously looked for this. While the practice of “salami slicing” is frowned upon, it is widely practised by those who think that the length of their curriculum vitae measures its strength, moreover it is inevitable in reporting major studies. The attempt to count total numbers of “articles” published also begs the question of what constitutes a scientific article. Even when this is resolved, classifying articles according to type remains problematic, one important issue the originators of the CONSORT guidelines identified was that not all randomised controlled trials (RCTs) could be identified as such solely on the basis of keywords.

In one sense, the authors are correct to regard the RCT as being at the top of the quality hierarchy, and consequently lament the continuing dearth of RCTs generally and in particular specialties. In another sense, this view is quite naive. RCTs and meta-analyses of RCTs are performed specifically to assess treatment effectiveness. There are many

other, equally valid study objectives. For instance, cohort and case-control studies, alluded to in the introduction, relate to elucidation of aetiology, which is essential from the standpoint of preventability of disease. If the medical literature consisted solely of RCTs, we would have to regard it as seriously deficient. In particular, we should be very apprehensive about using any treatment just because we can show it is effective, without understanding how it works.

In interpreting differences between specialties, the authors note that “surgeons may be apprehensive about randomising between two treatments, when they may have greater expertise and/or belief in the efficacy of one surgical treatment modality”. The point about belief in efficacy applies equally to non-surgical specialties and here there may be selection bias towards the authors’ own fields of expertise. Nevertheless, acquisition and demonstration of skill with an unfamiliar surgical technique is a real issue³ that may well have made RCTs comparing surgical techniques more difficult to perform than those in which treatment consists merely of popping pills. It may also reflect the disproportionately low research investment in surgical specialties compared with that made by the pharmaceutical industry. Moreover, the taxonomy of specialties used, while an obvious one, disregards the fact that medics are but one of the professional groups providing healthcare interventions for patients. For example, occasionally medical journals publish articles such as one showing that physiotherapy in the home or hospital setting is beneficial in multiple sclerosis.⁴ In contrast with medicine and dentistry, RCTs in the professions allied to medicine are in their infancy, and we should seek to encourage the rolling out of proper RCT methodology to these areas as well as those traditionally regarded as the preserve of doctors. There is an enormous difference between little evidence of efficacy and evidence of little efficacy. The fact that physiotherapy is to be tailored to the needs of the individual patient is not at

all a contraindication to RCT methodology but a challenge to adapt it in the most relevant manner.

The authors rightly draw attention to hindrances faced by research in general, and RCTs in particular. There has been a proliferation of hindrances—data protection and research governance regulations, and more recently the EU Directive concerning RCTs of medicinal substances—that has probably outweighed the beneficial impact of information technology on the ability to perform health related research. In recognition of the primary role of the RCT in determining the relative efficacy of different treatments, the CONSORT guidelines were developed, to raise the standard of reporting, and hence conduct, of RCTs specifically. In recommending the adoption of these guidelines to journals, one of us felt constrained to point out that the aim is not at all to place an additional hurdle that might deter researchers from doing RCTs.⁵

As well as a high journal birth rate there is also a death rate, particularly among journals that do not feature as part of Medline or Index Medicus (the latter point is all tied in with impact factors, etc). A Google search for references to journal death produced only one hit,⁶ João Magueijo’s article “Electronic Archives and the Death of Journals”, which presents a scenario of obsolescence of printed journals as their electronic counterparts take over, but does not include any quantitative data. In comparison with the massive research effort on extinction of animal species and death of lesser used languages, this admittedly less crucial issue seems to have attracted minimal attention—well worth someone somewhere doing a PhD on, not necessarily specifically in relation to healthcare. This may seem a trivial suggestion, with very little bearing on healthcare, nevertheless consider the crucial part that publication plays in the process that leads from healthcare research to the widespread implementation of its findings—it is important that as a profession we should understand all aspects of the publication process and its impact on dissemination and indeed research.

To come round full circle back to the article—the Portuguese name of the author alluded to above prompted one of us to recall the remark in Alexander McCall Smith’s novel *Portuguese Irregular Verbs*:⁷ “... one reviewer (of von Igelfeld’s treatise) had simply written, ‘There is nothing more to be said on the subject. Nothing.’ Von Igelfeld had taken this compliment in the spirit in which it had been intended, but there was in his view a great deal more to be said ...”. Our view is that it is rather a case of “... of making many books (and articles) there is no end” (Ecclesiastes 12: 12) – a millennia-old prophecy surpassed only by “... many shall run to and fro, and knowledge shall be increased” (Daniel 12: 4).

Postgrad Med J 2006;**82**:355–356.
doi: 10.1136/pgmj.2006.045138

Authors’ affiliations

R G Newcombe, Department of Epidemiology, Statistics and Public Health, University of Wales College of Medicine, Cardiff, UK

J Stebbing, The Chelsea and Westminster Hospital, London, UK

Correspondence to: Professor R G Newcombe, Department of Epidemiology, Statistics and Public Health, University of Wales College of Medicine, Cardiff CF14 4XN, UK; newcombe@cf.ac.uk

REFERENCES

- 1 **Gnanalingham MG**, Robinson SG, Hawley DP, *et al.* A 30 year perspective of the quality of evidence published in 25 clinical journals: Signs of change? *Postgrad Med J* 2006;**82**:397–9.
- 2 **Abbasi K**. Let’s dump impact factors! *BMJ* 2004;**329**:7471.
- 3 **Clarke D**, Newcombe RG, Mansel RE. The learning curve in sentinel node biopsy: the ALMANAC experience. *Ann Surg Oncol* 2004;**11**(suppl):211–155.
- 4 **Wiles CM**, Newcombe RG, Fuller KJ, *et al.* Controlled randomised crossover trial of physiotherapy on mobility in chronic multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2001;**70**:174–79.
- 5 **Newcombe RG**. Reporting of clinical trials in the IJ—the CONSORT guidelines. *International Endodontic Journal* 2004;**37**:1–2.
- 6 **Magueijo J**. Electronic archives and the death of journals. <http://www.ingenta.com/institute1999/magueijo.doc>.
- 7 **Smith AMcC**. *Portuguese irregular verbs*. Edinburgh: Polygon, 2003.