Citation Analysis in Health Care Sciences

Innovative Investigation or Seductive Pseudo-science?

M. Rigby
Keele University, Keele, Staffordshire, UK

1. Introduction – the Question

The paper by Hsu and Ho [1] is an elegant piece of research, assiduously undertaken. But is this an innovative piece of science, progressing further our understanding of how to evaluate the quality and value of scientific papers, institutions and authors, or is it a very elegant sand sculpture, impressive and convincing to look at, but with no robust substance?

This question on the value and meaning of analysis of bibliometrics could be posed in any domain, but it is particularly relevant in the health sciences sector where so many scientific measures abound yet fundamentals are often lacking, and it resonates with health informatics in particular where the exacting digital science is either a strong cornerstone of effective care or a pointless screen trace at considerable cost. We know only too well of remote vital sign monitoring systems which add no value as there is no emergency response structure, or studies of measures of gait which are valuable for children or athletes, but do nothing to assess whether the 70 year old lady with arthritis can walk to her local shop.

The underlying question is valid – how to measure the scientific value, or added value, of a paper, an author, or an institution? As with health care itself, for altruistic and societal reasons we want the optimal generation of published evidence – that is why creators of knowledge and authors of papers are in business, usually funded by public money. But as with health care, ‘good’ and ‘best’ have many meanings. Highly interventionist hospital procedures and patient-sensitive personalised home nursing support are very different essential partners in quality, but have different value in particular settings and conditions. Looking outside the health sector, for instance at hotels, what is ‘good’ for a business person needing space to work, wifi, and a restaurant in which to entertain is very different from what is ‘good’ for a pensioner coming to visit a friend in hospital, who needs unadorned comfort and economy. Both have certain common requirements – safety, warmth, cleanliness, and trustworthy service, but they are unlikely to share common definitions of ‘good’.

2. Analysis – the Problem

Yet the problem with assessing ‘good’ healthcare papers and institutions is even more challenging, as we do not have a third party assessment other than the peer review standard of the journal. Hence the initial attraction of looking to citations as the impartial measure of third party ‘consumer’ action – in effect a consumer response to the papers on offer, and akin to the money-follows-the-patient approach to hospital funding.
Hsu and Ho claim that "Highly cited articles, and most frequently cited articles, represent the most significant developments in a specific field. Highly cited articles provide a valuable insight into which authors, articles and topics are motivating the research field over time". However, their single cited evidence is from the field of veterinary medicine [2], which though complex is nevertheless much simpler than the totality of the healthcare field. Is there any robust evidence, particularly from healthcare, to show that such highly cited papers are really a fair reflection of motivation, greatest activity, or greatest yield of new knowledge?

3. The Citation Index Weakness

The citation index – the count of citations of a paper by subsequent authors, is fine as a broad brush – it shows which papers are used to develop the case of a subsequent paper; poor articles, and those which add nothing new, are least likely to be cited. But does repeated citation mean that a paper brings highly significant new knowledge, or has beneficial impact?

A full study of this question, across all the aspects of health care literature, and from different viewpoints of originality and utility, would be exhausting if it were to be exhaustive. But a quick dive into issues in child health and paediatrics severely challenges the hypothesis. For instance, in 1998 Wakefield and colleagues published in the Lancet a now infamous paper claiming a link between measles, mumps and rubella (MMR) vaccination and autism [3]. That paper, now disproved and withdrawn, has received over 100 PubMed citations in the subsequent 16 years. Seriously reduced subsequent immunisation levels as a result of the paper will have led to significant un-necessary morbidity and arguably mortality. By contrast, taking the example of sleeping position for infants, the benefits of putting an infant in the supine position to reduce the risk of cot deaths are well known, but the initiating paper is not – it is generally attributed to a study in 1985 in Hong Kong by Davies [4]. The subsequent 61% reduction in the incidence of cot deaths in the United States in the 12-year period 1992–2003 is available as a compilation graph of data from two US national sources displayed on Wikipedia [5] and obscurely attributed to the Safe to Sleep campaign of the National Institute of Child Health and Human Development [6], where finding the graph is a challenge failed by this author. Many other countries have similar national campaigns, many infant lives appear to have been saved – but where is the citation of, or attribution to, Davies? PubMed cites a total of 15 citations of his paper in the subsequent 30 years.

So in terms of effects on morbidity and mortality, Wakefield et al. had a negative effect and gain high citation rate status. Meanwhile Davies triggers a major life saving initiative from a single study and barely enters the citation radar. There is no attribution of the campaigns designed and the lives saved to the investment in funding of Davies and of his institution – the Department of Paediatrics of the Prince of Wales Hospital, Chinese University of Hong. Davies has even written a book on his own learning experiences of child health in Hong Kong [7], yet the significance of this investigative setting is not reflected in any metrics.

Hsu and Ho are extremely fair and open. They themselves outline in their section 4.4 [1] several significant weaknesses to citations rates, so strong that they could be argued to be fatal flaws when using citation indices as a pseudo-scientific measure as opposed to a broad brush picture.

3.1 Inaccuracy

In another scientific domain citation rates have been reported as having an error rate of 22% [8], and a quarter to be inappropriate in another domain [9]. There can surely be no place in health science for a measure which can be so inaccurate, nor any value in applying advanced analytic techniques to such unreliable data.

3.2 Credit for Reviewers

This same study found that in the field of ecology 15% of citations falsely attributed scientific ideas to reviewers of papers, rather than to the original authors of the initial work and resultant paper [8]. Is this source of error any less likely in the health and care fields, due not least to citation laziness, using an available citation rather than following good practice of finding the original before citing it? Reviews, or papers citing earlier original work, may indeed be cited more frequently that the original paper (as the Davies example above demonstrates vividly), but it is seldom the case that the review or discussion paper is individually more important than the original. Thus reviewers and commentators are potentially able to get more citations credited to them than creators of the original new knowledge.

3.3 Size of Field and Specialty Biases

As citations are a measure of how many times subsequent authors have referred to or built on prior published work, it is inevitable that there will be a bias in favour of the larger specialties – and the fact that in Hsu and Ho’s findings general medicine ranks highly and small specialties do not appear would support this [1]. But both in terms of personal achievement, and in terms of societal benefit, breakthroughs in understanding of a problem in neurology or virology could be even more valuable – they just lack the large scientific and clinical audiences. Secondly, the more laboratory based cell research, or pharmacochemical paper which is the foundation of a new treatment, will get less citations than the subsequent drug development and clinical trials papers – even where the trialists correctly cite their upstream enablers, it is the application papers that will receive the most citations, and there is no means within citation indices of apportioning derived credit to the initiating work and its publication. Thus it is easier to build up high citation rates in some specialties than others. Tainer points out that this bias not only works against such fundamental scientists, but that some journals fuel this by not publishing papers in low citation areas, even to the point of rejecting papers from such a field even though their content has major potential in enabling subsequent fruitful applied research [10].
3.4 Journal and Paper Type Biases

Publication policies of journals can themselves initiate biases acting against the deeper and more strongly reasoned and robust papers. One aspect is that mentioned above, the bias against low citation fields of study, and instead bias towards more cited fields which are more likely to boost the journal’s own impact factor. The second, also mentioned by Tainer [10], is the preference of some journals for short reports rather than full research papers. Such journals and short reports have a valid and important role in getting new knowledge to a larger audience and in an easily assimilated way, and may well cite the deeper sources and studies from which they draw. But as they will reach a wider audience they will get more citations than the more robust source studies; moreover some scientists will prefer for reasons of convenience only to publish in such sources. This produces a double bias in favour of the authors who favour short report formats, and against purer scientists, yet there is no regular means of weighting citation counts or indices to give fairer credit to the authors of occasional but foundational work. When citation indices start to be used as a measure of personal or institutional merit, or to be a basis for personal promotion or institutional funding without any study of content, this becomes unfair and also undesirable for reasons of scientific strength and balance.

3.5 Bias against Enablers and Team Members

Some scientific disciplines in the health domain are vital enablers to applied research and to developing treatment regimes. These enabling disciplines include statistics, and ethics, but others can be included such as psychology. Methodological and conceptual papers in these disciplines will be far fewer than their use in applied studies, which will receive the much more frequent citations. In many areas of health research, especially clinical trials and epidemiology, the presence of a statistician in the study team, and thus in the author team, is essential, but the statistician will very seldom be the lead or corresponding author. An ethicist, psychologist, nurse or other professional may well have had a key part in the design and success of the study – indeed, the lead author may only have a very marginally greater role than others, or may even be a titular team leader by way of medical position, but the enablers who have had a key role miss out in the citation index. The same may apply in purely academic teams, where some who receive extensive citations are more professional team builders and grant negotiators than active applied scientists – this is not in any way to denigrate those important roles and skills, but to emphasise that their second-author applied scientist workers are unfairly and inappropriately discriminated against in citation analyses.

3.6 Use of Tools and Paradigms

One regular reason for citations is to justify the use of a tool or assessment technique, whether clinical or methodological. But here too the playing field is not level. When a tool is used, it is normal to expect it to be validated – that means there should be reference to the original tool, and to the validation studies in a relevant field or domain. Can it be reliably believed that the citations in every paper, and thus in the overall citation indices, give appropriate credit to the original paper’s author as well as to the validators? The evidence of that happening currently seems lacking. Moreover, some tools and principles will get regularly cited, while other very influential phenomena attributable to a particular study and researcher are quite frequently mentioned without citation – the Hawthorne Effect is well-known but who has heard of Landsberger [11, 12]? Or that he only named it after the location of the studies thirty years after discovering the phenomenon?

3.7 Reasons for, and thus Meaning of, Citations

An initial assumption would be that citations are a reflection of the relevance of preceding theories or knowledge. But this is not the case. Citation may also be to criticise or refute – a citation count will not differentiate between the two. Indeed, one way to get highly cited could be to write provocatively. Hsu and Ho [1] quote in full Garfield’s list of 16 reasons for citing a paper, with there being nearly as many negative or critical reasons as there are laudatory ones [13] – though that is not to say that all are used equally. However, all sixteen reasons will be counted equally as citations.

In summary, therefore, citation indices can be seen as a very crude, and often potentially misleading, measure. It is rather like a parent feeling a child’s forehead with the back of their hand to see if the child is feverish – in an immediate context it is useful and generally indicative, but it is not scientific enough to be of use for clinical diagnosis, and moreover will fail if the child has just had a hot bath, or if the parent has just been out in sub-zero temperatures. But being negative is not enough on its own – can a constructive solution be found?

4. The Core Objectives

So what is a good measure? How do we assess utility and value? It is indeed tempting to use the available measures and move forward from these, as Hsu and Ho have done, but while this is spurred by the hope of better refining such measures it runs the risk of creating a spurious enhancement with little real world meaning. The alternative is to start with the vision – what do we really want to measure? – and then build back to establish data sources, direct or by proxy.

In health care sciences three objectives can be postulated:
1. Beneficially changing practice to produce better outcomes.
2. Improving and updating education.
3. Good value for money in innovative research and academic output.

4.1 Beneficially Changing Practice and Outcomes

This is the fundamental goal of health systems. Measurement of it, though, is a challenging matter. At the system level, not least with regard to aid programmes, there is some work, but at the intervention level this is more challenging.
There is a known problem of resistance and delay in getting published scientific results assimilated such that this is the subject of programmes such as Getting Research Into Policy and Practice (GRIPP) in many countries. In the midst of all this, citation back to the original author of an improvement is not the priority, and indeed given the important contributions of those who have validated, and refined, the original idea through follow-on studies, this is understandable as the desired progress in practice is the result of the dis-organised but enriching flow and fostering of ideas which is the nature and purpose of health research.

Were the line of measuring yield of practice benefit to be followed, the challenge of agreeing a common metrics system, such as Quality Adjusted Life Years (QALYs), and calculating everything to it attributed back to the initiator and the subsequent application research, would be enormous. Simpler metrics, such as citations in national policy documents, or those in professional and products of standards bodies, could have meaning when used, but would not be equal in value, and published papers that changed individual-level practice could pass unacknowledged. The conclusion must be that though assessment of impact of individual papers, discoveries, or resultant programmes of change is valid and important as a subject of study in its own right, there is no currently accessible uniform metric available to be utilised. Indeed, there is a real risk that any attempt that might be made would be more likely to find the populist issues rather than the kaizen effect of the more modest but widely useful and meaningful less eye-catching ideas, and at the same time to fail to reach the initial innovator.

4.2 Improving and Updating Education

All healthcare practice should be evidence based, and that evidence comes from the literature. However, there is no feedback to repositories or to authors as to when their material has been used quite legitimately and legally for teaching purposes, whether a book or a paper. Books have number of sales as a broad measure, but these are not collated other than by publishers, and there are no doubt considered commercial-in-confidence data. Searches for papers and other publications through professional social media sites such as Academia [14] and ResearchGate [15] will show the number of hits per paper (and per researcher) but will not be shared.

Thus the contributions of individual researchers, authors, or institutions to professional education in the health domain do not generate any routine measures which can be collated. This refers to clinical and healthcare research used in education, and not to published papers on healthcare education itself, which has its own bibliography. So as with improvement of care, for this use and impact it appears that there are not regular metrics capable of aggregate use and comparison.

4.3 Good Value for Money in Innovative Academic Output

This is the elusive target, shared by funders, public, and indeed the academic community. But there are two problems. The first is that often departments do not work in academic isolation, and may be linked to clinical or teaching activities – Davies, cited earlier [4], happens to be a case in point. So analysing costs is difficult, and the situation where organisations have to spend resources splitting staff time and overheads between functions is surely to be avoided as a general routine (though necessary if validating the costs linked to a specific funded study). Secondly, good departments should work in teams, not least to develop new researchers with potential, so some time should be spent on development and human investment, not on solely on outputs measurable in the short-term such as publications.

5. So What is the Answer?

It is good for academics and scientists, as much as clinicians, to practice reflectively and self-critically; it is good for funders and custodians of public funds to ensure that they are getting overall a good return on investment. But this needs to be in a wider and more mature context, and with a medium term horizon.

The health care comparison helps again. A ‘good’ clinician needs to have the relevant up-to-date evidence and guidance at his or her fingertips – but alone that is not enough. Relying solely on vital signs, protocols, and pharmacopeia will not make a good health professional – though these tools are vitally important. Two other things are needed – a wider viewpoint to see what else matters from the patient’s need and perspective; and the wisdom to know when the metrics are either giving a false reading or are not the only relevant factors.

Modern scientific clinical case management has two conflicting pressures – the legitimate desire to maximise knowledge and expertise, creating the risk of silo-based symptom care; and the desire to return to holistic integrated care based on Hippocratic principles [16, 17]. As each scientific strand is developed and refined, so it pulls against the core and initiating drive to give a balanced solution to the initial question.

The same would seem to apply to analysis of bibliometrics in the health sciences. Knowing that a paper is well cited is a measure of its value in some way, but to some degree this is also a matter of luck such as copycat citation or avoidance of a thorough search for a better reference. It is significantly removed from proof that is the most valuable or value adding paper in its field. The risk is that once a measure, such as automatic counting of citations as a feasible automated data source, becomes unduly depended on, sight is lost of the core objective of understanding the relative merits of different papers and sources, of which citation is only one aspect and value. And then there is the temptation to become even more sophisticated in the study of that one measure, and thus inflate it above its true value, and by implication imbue it with a falsely heightened scientific value. This is the route that Hsu and Ho seem to have followed.

As with most things in life, the answer seems to lie in pragmatism. First, to identify a number of different value measures which address different dimensions; then to apply these only occasionally as the
more valuable ones may be more difficult and expensive, and at a long enough interval to make gaming not worthwhile; and finally to apply these with insight rather than by formula. Investing in sophisticated methodologies for any one indicator, or creating a new monitoring industry, are not likely to yield adequate societal returns on investment, and will do nothing to stimulate innovative thinking, new insights or fair allocation of resources or personal promotion to those who produce original thought.

References