Assessing non-standard article impact using F1000 labels[1]

Ehsan Mohammadi, Mike Thelwall
Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.
E-mail: e.mohammadi@wlv.ac.uk, m.thelwall@wlv.ac.uk

Abstract

Faculty of 1000 (F1000) is a post-publishing peer review web site where experts evaluate and rate biomedical publications. F1000 reviewers also assign labels to each paper from a standard list or article types. This research examines the relationship between article types, citation counts and F1000 article factors (FFa). For this purpose, a random sample of F1000 medical articles from the years 2007 and 2008 were studied. In seven out of the nine cases, there were no significant differences between the article types in terms of citation counts and FFa scores. Nevertheless, citation counts and FFa scores were significantly different for two article types: "New finding" and "Changes clinical practice": FFa scores value the appropriateness of medical research for clinical practice and "New finding" articles are more highly cited. It seems that highlighting key features of medical articles alongside ratings by Faculty members of F1000 could help to reveal the hidden value of some medical papers.

Keywords: Faculty of F1000, Altmetrics, Beyond impact, Research assessment, Post-publishing peer review

Introduction

The medical research community and funders are seeking ways to examine the quality of research outputs from different points of view (Chalmers & Glasziou, 2009). Research impact in medical sciences is defined as "any type of output of research activities which can be considered a positive return for the scientific community, health systems, patients, and the society in general" (Banzi, Moja, Pistotti, Facchini, & Liberati, 2011). Although citation analysis has been employed widely for research evaluation, it has some limitations (MacRoberts & MacRoberts, 1996). Moreover, citation analysis is able to capture research influence in publications from the knowledge advancement perspective only (Kostoff, 1998) and so more aspects should be considered for measuring the different research impacts of medical publications.

Knowledge transfer from research to the clinic, the implementation of research findings in practice (Sarli, Dubinsky, & Holmes, 2010), is critical for much medical research but the best method for measuring this is unknown (Weiss, 2007). One investigation studied references from the UK cancer clinical guidelines to capture knowledge transfer from research to practice, finding that most references were to journal articles from the subfield (Lewison & Sullivan, 2008).Moreover, new drug applications are an alternative way to assess pharmaceutical research output (Koenig, 1982) but supplementary ways are needed to examine medical research from the drug discovery perspective. Although it is not easy to measure the impact of research on teaching (Zaman & Britain, 2004), some studies have used online syllabuses (Kousha & Thelwall, 2008), web citations (Vaughan & Shaw, 2005) and medical textbooks (Lewison, 2005) to measure the influence of research on educational activities. These days, with the sheer amount of scholarly information a systematic method is needed to highlight the appropriateness of medical publications for teaching.

Another problem is that citations can be negative and hence citation counts can sometimes be misleading (Small, 2004). Disagreement in scholarly communications is useful (Harnad, 1985) especially in medical sciences where there are a number of poor-quality papers (Cole, Cole, & Simon, 1981).

Faculty of 1000 (F1000) is a post-publication peer review system for the evaluation of biomedical journal articles that partly exists because of the limitations of citation analysis. Around 10,000 researchers and clinicians across the world evaluate papers in more than 40 subjects of the biomedical domain (F1000, 2012a). In addition to rating research papers, reviewers tag articles with predefined labels, including appropriateness for changes to clinical

practice, suitability for new drugs, and usability for teaching. They also classify whether new papers confirm, challenge, and reject hypotheses from previous research. F1000 reviewers try to add value to their evaluations by distinguishing key features of biomedical papers that could represent different dimensions of research. This study examines the relationship between highlighted features, assigned labels, of medical articles by faculty member of F1000 and their FFa scores as well as citation counts in order to assess whether FFa scores can help to identify types of article that are successful but not highly cited.

Related research

Research assessment in medical science

Several frameworks have been designed to evaluate medical research impact. Kuruvilla and colleagues (2006) used four main categories including research-related impacts, policy impacts, service impacts and societal impacts. Weiss (2007) criticised medical research evaluation systems as being output-based and measuring only productivity. He therefore suggested an outcome-oriented model for measuring medical research instead. Tomlinson (2000) investigated the impact of the 2001 UK Research Assessment Exercise (RAE) of medical sciences. He believed that although the exercise led to an increase in research, it underestimated clinical and health approaches. A team from the Becker Medical Library proposed a practical model for the evaluation of medical research impact beyond traditional citation analysis (Sarli, Dubinsky et al. 2010) including research output, knowledge transfer, clinical, implementation, community benefit, legislation and policy enactment. Recently, Banzi and colleagues (2011) reviewed research assessment of medical sciences literature and defined the main research impacts as “advancing knowledge”, “capacity building”, “informing decision-making”, “health benefits”, and “broad socio-economic benefits”. It can be concluded that policy makers, researchers and funders try to optimize the current research assessment methods to measure different and broader impacts of medical research.

Some studies have focused on measuring research impact outside of the research community. An investigation into clinical publications about nursing found evidence of knowledge transfer from researchers to clinicians in the field (Niederkrotenthaler, Dorner et al. 2011). Moreover, citation analysis has been used to capture research influence in government policy documents, textbooks and clinical guidelines (Lewison, 2005). In another study, the influence of articles for The New England Journal of Medicine on patients was

measured and a positive impact on patients was reported (Price and Simon, 2009). Additionally, a "societal impact factor" has been suggested as a complementary measure for the journal impact factor (Niederkrotenthaler, Dorner et al. 2011). Perhaps related to this, the impact of medical research on mass media was examined and the BBC was found to have cited around 1380 research articles related to cancer between1998 and 2006 (Lewison, Tootell et al. 2008).

Motivations for citing

Since researchers cannot cite all previous articles, they need a strategy to select their references. Two main theories related to citation behaviour, normative and social constructivist, have been developed during the past decades. The former theory assumes that scholars give credit to previous publications via citations (Bornmann & Daniel, 2008) whilst the latter does not. Camacho-Miñano and Núñez-Nicke (2009) suggested that a multi-layered model is necessary to explain citation practices better. Although Cronin (1984) and MacRoberts and MacRoberts (2010) doubt that precise reasons for citation behavior can be uncovered, several empirical studies have been nonetheless carried out to discover motivations for citation (See Bornmann & Daniel, 2008). In the medical sciences, an investigation into uncited biomedical articles revealed that the number of authors and references in cited papers was significantly greater than in uncited articles (Stern, 1990). In another study, a combination of different variables including locations and reasons for each citation as well as the type of research (basic or clinical) being discussed were used to assess the outcomes of medical publications. Despite the importance of surgery articles, they were cited less in the literature. As a result, citation counts cannot be used as the only indicator for research assessment in medical sciences (Hanney et al., 2005). Furthermore, features of highly cited and rarely cited articles of The Lancet were explored based on variables such as the number of authors, references and citations plus other characteristics related to article type. Articles with more authors and abstract words were cited dramatically more than the lowest cited papers (Kostoff, 2007). Recently, Jones and colleagues (2012) conducted a literature review to extract objective and subjective features of citations to develop a new method for citation categorization to assess the quality of individual articles. They claimed that the results of their survey could be useful to measure the impact of biomedical publications in different generations based on citing publications.

Original research data from an investigation can be reused in new research (Fienberg & Martin, 1985) but some articles with original data have not been cited or cited rarely although their data has been used. This means that some important articles in terms of original data could not be recognized through citation analysis (MacRoberts & MacRoberts, 2010) whereas the data set could be considered as a unit for measuring research impact (Sarli & Holmes, 2012) in medical publications. Results of an investigation on cancer microarray indicated that clinical trials that published their data publicly were cited 70% more than those that did not. This suggests that the availability of research data is an important reason for citation (Piwowar, Day, & Fridsma, 2007) in the medical sciences.

Expert judgments versus bibliometric measures

Expert judgement is often used as the primary method to evaluate research and its results are sometimes compared to bibliometric measures to assess the value of the latter. Maier (2006) used expert opinions and the journal impact factors for rating top journals in the field of regional science, finding no significant positive association between the two measures. Peer judgements and the journal impact factors of research articles by Italian researchers in three fields (chemistry, biology, and economics) have also been compared, finding a significant but not strong correlation between the two variables (Reale, Barbara, & Costantini, 2007). Moreover, journal impact factors and article citations were compared to peer judgments in an Italian national research evaluation program, findings a significant correlation between both bibliometric indicators and expert ratings (Franceschet & Costantini, 2011).

Several investigations have compared bibliometric indictors and the results of the UK research assessment exercises (UK RAE), an expert-based method of research evaluation. Oppenheim (1995) compared citation counts and the results of the 1992 UK RAE for library and information science departments and found a significant and strong correlation between the two measures. Similarly, a strong significant correlation was observed between UK RAE scores and several bibliometric indicators including total citation, citation count per publication, and citation count per academic staff for British library and information science departments (Seng & Willett, 1995). Additionally, Smith & Eysenck (2002) investigated the association between citation counts and the UK RAE scores in 1996 and 2001 for the British psychology departments and a strong significant correlation between the two indicators was reported.

Norris and Oppenheim (2003) compared different citation metrics such as total and average researchers' citations with the results of the 2001 UK RAE scores for archaeology. The results showed high correlations between citation indicators and UK RAE thus they suggested that citation measures may be a useful source for initial rating of research departments. Furthermore, comparison of the 2001 UK RAE and citation measures in the field of music exhibited a strong correlation between the indicators at the level of department while the correlation was weaker for individual citation counts. They suggested that more types of citation data beyond journal articles should be considered for evaluation of music literature (Oppenheim & Summers, 2008). A large-scale study investigated correlations between citations and UK RAE 2001 scores across a variety of disciplines. In some fields of social sciences including education, history, sociology, social policy and administration, politics and international studies there was no correlation between the indicators (Mahdi, D'Este, & Neely, 2008). Recently, Kousha, Thelwall, and Rezaie, (2011) investigated citations of books submitted to the 2008 UK RAE for seven disciplines of social sciences and humanities. They found that there is enough citation data for the book-based disciplines that would be useful for the expert-oriented research assessment exercise.

In an old study, the research performances of six research groups were measured with both peer judgements and citation analysis and the results have also been compared. Although each method had its own unique features to assess research performance, their results of both correlated (Nederhof & Van Raan, 1993). The relationships between various citation-based indicators and expert ratings for different research groups for a university in Norway were measured. There was a positive but not strong correlation between expert ratings and all the bibliometric indicators. The authors argued that citation metrics could not be an absolute alternative for peer evaluation but only a complimentary indicator (Aksnes & Taxt, 2004). Van Raan (2006) compared the h-index and the Crown indicator (CPP/ FCSm[2]), both citation-based, with expert evaluations of chemistry research groups in the Netherlands. The quantitative and qualitative indicators associated very well but the new Crown indicator was more suitable for smaller research groups with "less heavy citation traffic". Opthof and

[2] The CPP/FCSm is an indicator which developed by Centre for Science and Technology Studies (CWTS) with the aim of normalization of citation among different fields. It has been renamed as crown indicator (See http://arxiv.org/pdf/1003.2167.pdf).

Leydesdorff (2011) challenged the new crown indicator in a study. They used previously-released data (Van Raan, 2006) to re-examine correlations between the citation parameters and peer review results for chemistry research groups of the Netherlands. The qualitative indicator and the two proposed citation metrics by the CWTS (CPP/ FCSm) had no significant correlation. Waltman and colleagues (2011) believed that Opthof and Leydesdorff (2011) used a statistical method that only shows "presence and the absence of a relation" rather than determining the degree of correlation between the two variables. Thus, Waltman and colleagues (2011) investigated a larger amount of data with a more suitable statistical technique to measure degree of relation between the two indicators and revealed that the CPP/ FCSm parameters correlated significantly with expert opinions.

F1000 and research evaluation

Reviewers of F1000 now apparently include 10,000 experts in 44 subject areas of medical sciences and biology (F1000, 2012b). The method of assessment in F1000 is based on the recommendations of reviewers (Wets, Weedon, & Velterop, 2003) regardless of the impact factor of the publishing journal.

F1000 evaluates over 1500 articles monthly from 3500 different publications (see http://f1000.com/about/whatis/coverage) but the list changes constantly. The most popular and high prestige journals of the disciplines related to biology and medicine, such as Nature, Science, Cell, New England Journal of Medicine, and Journal of Experimental Medicine are covered by F1000. However, 85% of the evaluated articles are selected from less high profile journals. F1000 sends the table of contents of relevant journals to Associate Faculty Members each month. Each associate then scans one general and one special journal based on their speciality and coordinates with other associates to select articles for evaluation (F1000, 2012a). F1000 associates review and rate articles according to three positive levels: Exceptional, Must Read, and Recommended. The evaluation system has no option for negative or neutral assessments. Each evaluated article can be reviewed several times. The F1000 Article Factor (FFa) is based on all reviewers' scores. FFa is calculated based on the highest score assigned by reviewers - Exceptional, Must Read, and Recommended valued at 10, 8, and 6 respectively - and then adding an incremental value for each additional rating (3 for Exceptional, 2 for Must Read, 1 for Recommended) (Huggett, 2012). F1000 members also classify the assessed articles into several categorises based on article type (F1000, 2012a)

A few studies have used F1000 in research evaluation. The Wellcome Trust, a research-funding agency in the UK, used their expert reviewers' assessments as well as F1000 article ratings and compared them with citation counts for their own funded research. Although only 7% (n=48) of their studied articles had been reviewed by F1000, there was a positive correlation between Wellcome Trust reviewers' assessments and F1000 ratings. Nevertheless, some publications which were marked by the two groups of reviewers as highly rated were not highly cited three years after the date of publication (Allen, Jones, Dolby, Lynn, & Walport, 2009). In another investigation, Wardle (2010) compared F1000 rates and citations for the articles of seven important ecological journals. The results showed that 46% and 31% of all publications which were highlighted by reviewers of F1000 had not been highly cited articles.

Li and Thelwall (2012) compared FFa score of 1,397 reviewed articles in the field of Genomics & Genetics with citations and the journal impact factor. They found significant correlations between FFa scores and both citation metrics. Recently, comparing FFa score and several bibliometrics measures provided by InCitesTM Thomson Reuter, Bornmann and Leydesdorff (2012) found the highest correlation between the FFa score and "Percentile in Subject Area" amongst others.

Research questions

The real impact of medical research cannot be captured by current methods and so different approaches are needed to measure the true impact of publications in this area. The "altmetrics" movement is a new approach to measure invisible and broader research impact in a "diverse scholarly ecosystem" beyond classical measures (Priem, Taraborelli, Groth, & Neylon, 2011). F1000 has been suggested as one of altmetrics resources that is powered by crowdsourcing peer-review (Priem & Hemminger, 2010).

Some studies have used the FFa score and found a good correlation between this measure and citation-related metrics (Bornmann & Leydesdorff, 2012; Li & Thelwall, 2012) but this research examines the post-publishing peer review system from a new perspective. Labels assigned to reviewed articles by F1000's reviewers and their relationships with citations and FFa scores are investigated. The questions below drive this study:

1. Does the type of an article impact on its F1000 rating?
2. Does the type of an article impact on its citation count?

Data collection and methodology

F1000 has been selected as the primary source of data collection because the main aim of this research is to examine the relationship between the highlighted features of reviewed articles, citations and FFa scores. Our sample is limited to medical articles because the predefined classifications for articles are more suitable for medical research than biology. The peak time for receiving citations is normally three years after a journal article is published (Moed, 2005); therefore, in this research the time span was restricted to articles published in 2007 and 2008. By the time of the data collection, January 2012, 3307 and 5091, articles published in 2007 and 2008, respectively, had been evaluated by F1000 medicine. A random sample of 350 out of 3307 and 550 out of 5091 records for 2007 and 2008 was selected. In the next step, FFa scores and labels of sample articles were extracted manually from F1000. The available labels at the time of data gathering were used, although they are occasionally changed by F1000.

Since previous research has found a strong association between WoS and Scopus, the two main citation sources (Archambault, Campbell, Gingras, & Lariviere, 2009), Scopus was selected for citation data as its coverage is greater than WoS (Falagas, Pitsouni, Malietzis, & Pappas, 2008). Thus, citation data for each article in our sample was downloaded from Scopus with manual searching based on the article title in quotation marks in the article title drop-down menu. Some records were removed as they were duplicated in Scopus with different citation counts, leaving 344 and 533 sampled articles from 2007 and 2008 for this study.

In order to investigate relationships between the main features of articles with FFa scores and citations, we used the non-parametric Mann-Whitney U-test since citations and FFa scores were skewed for the sample of data.

Results

There is a low but significant correlation between FFa scores and citations for both 2007 (Spearman r=0.383, n=344, p=0.01) and 2008 (Spearman r=0.300, n=533, p=0.01).

As Table 1 shows, most of the sampled articles were evaluated only once by F1000 faculty. There was a strong correlation (Spearman r=0.509, p=0.01) between evaluation frequencies and FFa scores for all studied articles, which is unsurprising since evaluation frequency contributes to overall FFa scores.

Table1. Frequency of F1000 evaluations for the sampled reviewed articles in 2007 and 2008.

Evaluation frequency	1	2	3	4	6	8
Number of articles	89.5%	8.5%	1.5%	0.3%	0.1%	0.1%

Table 2 compares the decisions made by different reviewers for the studied articles. The reviewers assigned the lowest level of evaluation (Recommended) to the majority of the investigated papers (65.8%) while only 4.8% were assigned the highest level (Exceptional).

Table 2. F1000 articles based on evaluation times and ratings for articles sampled from 2007 and 2008.

No of Reviews	1	2	3	4	6	8	Total
Articles	784	74	14	3	1	1	877
Exceptional	3%	1%	0.5%	0.1%	0.1%	0.1%	4.8%
Must Read	20%	6.2%	2%	0.3%	0.5%	0.3%	29.4%
Recommended	55.5%	7.6%	1.7%	0.8%	0	4(0.4%)	65.8%

A majority of 2007 (72%, n = 247) and of 2008 (80%, n = 441) articles were reviewed in the publication year .

Around 94% (n=822) of the articles have been labeled at least once with one of the pre-defined categories. Around 56% (n=450) of the articles received one label, while 35% (n=285) were tagged twice, and 11% were labeled three or more times in different categories. There was a significant but not strong positive correlation between the number of assigned labels and FFa scores (Spearman r=0.242, p=0.01) as well as the number of labels and citation counts (Spearman r=0.201, p=0.01).

As shown in Table 3, the majority of the tagged articles were labeled “New Finding” (54%) or “Confirmation” (43%) while 25% of those articles were categorized as “Clinical Trial”. Additionally, 11% and 10% of those articles were tagged “Controversial” and “Technical Advance” respectively. “Changes Clinical Practice” and “Review” were assigned to only 8% and 6% of the papers. Few articles were labeled “Refutation” (1.5%) or “Novel Drug Target” (.5%).

In order to control Type I error, a Bonferroni correction was used for the Mann–Whitney U tests. So, rather than using .05 as critical level of significance, we used 05/9=0.005 (level of significance/ number of tests). Exact significances should be used in Mann–Whitney U tests when sample sizes are small (Field, 2009), so in this study for the "Refutation" and "Novel Drug Target" categories exact significance was considered.

The differences between the assigned labels for both FFa scores and citations in most cases were not statistically significant. Table 3 shows that both FFa scores and citations of articles which were labeled as "Confirmation", "Clinical Trial", "Controversial", "Technical advance", Review", "Refutation" and "Novel Drug Target" were not significantly different from those articles which were not categorized in these classifications (i.e., an indicative p value <0.005). In contrast, there were significant differences in FFa scores for articles which labeled as "Changes to clinical practice" and papers that were not classified in this category. Furthermore, articles labeled as "New Finding" also show significant differences from articles that were not classified under this label in terms of citation counts. The median number of citations to articles classified as "New Finding" (34) was more than for papers that were not categorized (23) under this article type. The median of FFa scores of articles were categorised as "Changes to clinical practice" and other papers was similar (8).

Table3. Associations between labels assigned to F1000 articles in 2007 and 2008 and FFa scores and citation counts.

Labels	FFa Scores			Citation Counts	
	N	Mean Rank	P-value	Mean Rank	P-value
New Finding	438	428.34	0.012	447.04	0.000
Not New Finding	384	392.29		370.97	
Confirmation	351	430.16	0.024	407.61	0.685
Not Confirmation	471	397.59		414.4	
Clinical Trial	206	433.4	0.075	445.08	0.019
Not Clinical Trial	616	404.18		400.27	
Controversial	93	410.45	0.958	437.55	0.261

Not Controversial	729	411.63		408.18	
Technical Advance	81	432.4	0.332	399.15	0.622
Not Technical Advance	741	409.22		412.85	
Changes to Clinical Practice	62	572.92	0.000	480.34	0.018
Not Changes to Clinical Practice	760	398.33		405.88	
Review	50	461.79	0.072	485.56	0.023
Not Review	772	408.24		406.7	
Refutation	12	497.88	0.152	591.75	0.007
Not Refutation	810	410.22		408.83	
Novel Drug Target	3	383.67	0.898	487.17	0.600
Not Novel Drug Target	819	411.6		411.22	

Discussion

The significant difference between the citation counts of articles which were classified as "New Finding" and other articles suggests that the 54% of articles with novel findings tend to receive more citations than other articles, presumably because the new findings are more useful for future published research, even though there was no evidence that such articles were more highly rated by the F1000 system. Moreover, there was a significant difference between the FFa scores of the 8% papers which were classified as "Changes Clinical Practice" and other articles, perhaps because the appropriateness of medical research for clinical practice is a feature of medical articles that is highly valued by experts even though it may not lead to increased citations. It seems that the FFa score is able to recognize appropriate articles for clinical practice better than citations and this is logical because citation practice is restricted to authors' activities while the suitability of an article for clinical section should be investigated from the practitioners' points of view.

The current investigation has several limitations. Each article could be assigned to more than one discipline in F1000. As a result, the sum of the papers of all the individual

disciplines of medical sciences (68,627) was almost twice as many as the total number of F1000 medicine articles (35,232). On average, an article was assigned to two disciplines in F1000. Moreover, an article can be appointed to a discipline in F1000 even it has not been reviewed by faculty members from the category. Thus, defining a subject for a reviewed article, particularly for multidisciplinary papers, is difficult. As mentioned before, the sample of this study was selected from the whole of F1000 medicine. Although the average "field citedness" is similar in some clinical medical disciplines (Seglen, 1997) it is not similar for all sub-fields of medical sciences (Harnad, 1985). Therefore, the citation propensity for the investigated sample from all medical disciplines can affect the results of the present research. Additionally, a paper has many chances to be reviewed by faculty members of several disciplines and the final FFa score is calculated based on all individual ratings of experts across different disciplines. This means that the papers which were assigned to a discipline in F1000 don't reflect only the activities of the faculty members of the category. F1000 reviewers and the number of reviewed papers were not distributed uniformly across the disciplines (see appendix 1). The academic position of F1000 experts is unknown despite the fact that the academic status of reviewers is an important issue (Zuccala, 2010) in a peer review system. Finally, the points scheme used by F1000 to convert judgements into a score is relatively arbitrary and a different scheme would produce a different ranking order of the articles. Nevertheless, any change in the points scheme would be unlikely to affect the restuls of the study here because the majority of the studied articles were only reviewed once (89.4%) and for these articles the actual scores assigned are irrelevant for their ranking (and hence for the tests used here). As a result, a different scoring system would make little difference to the results.

Overall, the assigned labels alongside the peer review rating show that some types of research can be identified by reviewers as being, on average, more valuable than others even if this will not be recognised by citations. Hence F1000 could be used in research evaluation exercises when the importance of practical findings needs to be recognised. Furthermore, since the majority of the studied articles were reviewed in the publication year, F1000 could be useful when quick evaluations are needed.

Appendix 1: Faculty members of F1000 in different disciplines of medical sciences

Topic	Associated and faculty members	Articles associated with the topic	Articles reviewed per faculty member
Anesthesiology & Pain Management	447	4803	11
Cardiovascular Disorders	156	3787	24
Critical Care & Emergency Medicine	141	3712	26
Dermatology	221	2591	12
Diabetes & Endocrinology	229	3744	16
Gastroenterology & Hepatology	296	3869	13
Hematology	238	2633	11
Infectious Diseases	253	4962	20
Nephrology	149	2002	13
Neurological Disorders	413	6183	15
Oncology	172	5371	31
Ophthalmology	175	536	3
Otolaryngology	149	1479	10
Psychiatry	180	3544	20
Public Health & Epidemiology	100	5314	53
Research Methodology	41	690	17
Respiratory Disorders	231	3520	15
Rheumatology & Clinical Immunology	313	3259	10
Urology	152	2397	16
Women's Health	124	4231	34
Total	4180	68627	

References

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. Research Evaluation, 13(1), 33–41.

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. PloS one, 4(6), e5910.

Archambault, E., Campbell, D., Gingras, Y., & Lariviere, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7).

Banzi, R., Moja, L., Pistotti, V., Facchini, A., & Liberati, A. (2011). Conceptual frameworks and empirical approaches used to assess the impact of health research: an overview of reviews. Health research policy and systems / BioMed Central, 9, 26. doi:10.1186/1478-4505-9-26

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64(1), 45–80.

Bornmann, L., & Leydesdorff, L. (2012). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. Digital Libraries; Applications. Retrieved from http://arxiv.org/abs/1211.1154

Camacho-Miñano, M.-M., & Núñez-Nickel, Manuel. (2009). The multilayered nature of reference selection. Journal of the American Society for Information Science and Technology, 60(4), 754–777. doi:10.1002/asi.21018

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. Lancet, 374(9683), 86–9. doi:10.1016/S0140-6736(09)60329-9

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. SCIENCE, 214(4523), 881–886.

Cronin, B. (1984). The citation process. The role and significance of citations in scientific communication. London: Taylor Graham.

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. Journal of Information Science, 27(1), 1–7.

F1000. (2012a). About F1000. Retrieved from http://f1000.com/prime/about/whatis

F1000. (2012b). F1000 Faculty. Retrieved from http://f1000.com/prime/thefaculty

Fienberg, S. E., & Martin, M. E. (1985). Sharing research data. Natl Academy Pr.

Franceschet, M., & Costantini, A. (2011). The first Italian research assessment exercise: a bibliometric perspective. Journal of Informetrics.

Hanney, S., Frame, I., Grant, J., Buxton, M., Young, T., & Lewison, G. (2005). Using categorisations of citations when assessing the outcomes from health research. Scientometrics, 65(3), 357–379. doi:10.1007/s11192-005-0279-y

Huggett, S. (2012). F1000 Journal Rankings: an alternative way to evaluate the scientific impact of scholarly communications. Research Trends, (26).

Jones, T. H., Donovan, C., & Hanney, S. (2012). Tracing the wider impacts of biomedical research: a literature search to develop a novel citation categorisation technique. Scientometrics, 1–10.

Koenig, M. E. D. (1982). Determinants of expert judgement of research performance. Scientometrics, 4(5), 361–378. doi:10.1007/BF02135122

Kostoff, R. n. (1998). The use and misuse of citation analysis in research evaluation. Scientometrics, 43(1), 27–43. doi:10.1007/BF02458392

Kostoff, R. n. (2007). The difference between highly and poorly cited medical articles in the journal Lancet. Scientometrics, 72(3), 513–520. doi:10.1007/s11192-007-1573-7

Kousha, K., & Thelwall, M. (2008). Assessing the impact of disciplinary research on teaching: An automatic analysis of online syllabuses. Journal of the American Society for Information Science and Technology, 59(13), 2060–2069.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. Journal of the American Society for Information Science and Technology, 62(11), 2147–2164. doi:10.1002/asi.21608

Kuruvilla, S., Mays, N., Pleasant, A., & Walt, G. (2006). Describing the impact of health research: a Research Impact Framework. BMC Health Services Research, 6(1), 134.

Lewison, G. (2005). Citations to papers from other documents. Handbook of Quantitative Science and Technology …. Retrieved from http://www.springerlink.com/index/T2H0245570526217.pdf

Lewison, G., & Sullivan, R. (2008). The impact of cancer research: how publications influence UK cancer clinical guidelines. British Journal of Cancer, 98(12), 1944–1950. doi:10.1038/sj.bjc.6604405

Li, X., & Thelwall, M. (2012). F1000 , Mendeley and Traditional Bibliometric Indicators. 17th International Conference on Science and Technology Indicators (Vol. 3, pp. 1–11).

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. Scientometrics, 36(3), 435–444.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. Journal of the American Society for Information Science and Technology, 61(1), 1–12.

Mahdi, S., D'Este, P., & Neely, A. D. (2008). Citation counts: are they good predictors of RAE scores?: a bibliometric analysis of RAE 2001. AIM Research.

Maier, G. (2006). Impact factors and peer judgment: The case of regional science journals. Scientometrics, 69(3), 651–667.

Marjanovic, S., Hanney, S., & Wooding, S. (2009). A historical reflection on research evaluation studies, their recurrent themes and challenges. technical report. RAND Corporation.

Moed, H. F. (2005). Citation analysis in research evaluation (Vol. 9). Kluwer Academic Pub.

Nederhof, A. J., & Van Raan, A. F. J. (1993). A bibliometric analysis of six economics research groups: A comparison with peer review. Research Policy, 22(4), 353–368.

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. Journal of Documentation, 59(6), 709–730.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British library and information science university departments. Journal of Documentation, 51(1), 18–27.

Oppenheim, C., & Summers, M. A. C. (2008). Citation counts and the Research Assessment Exercise, part VI: Unit of assessment 67 (music). Information Research, 13(2), 3.

Opthof, T., & Leydesdorff, L. (2011). A comment to the paper by Waltman et al., Scientometrics, 87, 467–481, 2011. Scientometrics, 88(3), 1011–1016.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. PloS one, 2(3), e308. doi:10.1371/journal.pone.0000308

Priem, J., & Hemminger, B. M. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. First Monday, 15(7), http://frodo.lib.uic.edu/ojsjournals/index.php/fm/. Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). alt-metrics: A manifesto (v 1.01 – September 28, 2011: removed dash in alt-metrics). Web.< http://altmetrics.org/manifesto.

Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: lessons from the Italian experience. Research Evaluation, 16(3), 216–228.

Sarli, C. C., Dubinsky, E. K., & Holmes, K. L. (2010). Beyond citation analysis: a model for assessment of research impact. Journal of the Medical Library Association: JMLA, 98(1), 17–23.

Sarli, C. C., & Holmes, K. L. (2012). The Becker Medical Library Model for Assessment of Research Impact. Bernard Becker Medical Library, Washington University School of Medicine.

Seng, L. B., & Willett, P. (1995). The citedness of publications by United Kingdom library schools. Journal of Information Science, 21(1), 68–71.

Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. Scientometrics, 60(1), 71–79. Retrieved from http://www.springerlink.com/index/X6VTVM1209131570.pdf

Smith, A. T., & Eysenck, M. (2002). The correlation between RAE ratings and citation counts in psychology.

Stern, R. E. (1990). Uncitedness in the biomedical literature. Journal of the American society for information science, 41(3), 193–196.

Tomlinson, S. (2000). The research assessment exercise and medical research. British Medical Journal, 320(7235), 636–639.

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. Scientometrics, 67(3), 491–502.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. Journal of the American Society for Information Science and Technology, 56(10), 1075–1087.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). On the correlation between bibliometric indicators and peer review: reply to Opthof and Leydesdorff. Scientometrics, 1–6.

Wardle, D. A. (2010). Do'Faculty of 1000'(F1000) ratings of ecological publications serve as reasonable predictors of their future impact? Ideas in Ecology and Evolution, 3(0).

Weiss, A. P. (2007). Measuring the impact of medical research: moving from outputs to outcomes. American Journal of Psychiatry, 164(2), 206.

Wets, K., Weedon, D., & Velterop, J. (2003). Post-publication filtering and evaluation: Faculty of 1000. Learned Publishing, 16(4), 249–258. doi:10.1087/095315103322421982

Zaman, M. uz, & Britain, G. (2004). Review of the Academic Evidence on the Relationship between Teaching and Research in Higher Education. Retrieved from https://www.education.gov.uk/publications/eOrderingDownload/RR506.pdf