

Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study

Cynthia Lokker, K Ann McKibbin, R James McKinlay, Nancy L Wilczynski, R Brian Haynes

Health Information Research Unit,
Department of Clinical
Epidemiology and Biostatistics,
McMaster University Faculty of
Health Sciences, Hamilton, ON,
Canada L8N 3Z5

Correspondence to: C Lokker
lokkerc@mcmaster.ca

BMJ 2008;336:655-7

doi:10.1136/bmj.39482.526713.BE

ABSTRACT

Objective To determine if citation counts at two years could be predicted for clinical articles that pass basic criteria for critical appraisal using data within three weeks of publication from external sources and an online article rating service.

Design Retrospective cohort study.

Setting Online rating service, Canada.

Participants 1274 articles from 105 journals published from January to June 2005, randomly divided into a 60:40 split to provide derivation and validation datasets.

Main outcome measures 20 article and journal features, including ratings of clinical relevance and newsworthiness, routinely collected by the McMaster online rating of evidence system, compared with citation counts at two years.

Results The derivation regression equation accounted for 60% of the variation ($R^2=0.60$, 95% confidence interval 0.538 to 0.629). This model applied to the validation dataset gave a similar prediction ($R^2=0.56$, 0.476 to 0.596, shrinkage 0.04; shrinkage measures how well the derived equation matches data from the validation dataset). Cited articles in the top half and top third were predicted with 83% and 61% sensitivity and 72% and 82% specificity. Higher citations were predicted by indexing in numerous databases; number of authors; abstraction in synoptic journals; clinical relevance scores; number of cited references; and original, multicentred, and therapy articles from journals with a greater proportion of articles abstracted.

Conclusion Citation counts can be reliably predicted at two years using data within three weeks of publication.

INTRODUCTION

The number of times an article is cited is an attractive measure of importance but it has no applicability until citation counts accrue, peaking at around three years.¹

Several attributes are associated with higher citation counts, such as nationality,^{2,3} number of authors,^{4,5} number of pages,⁴ and online (open access) availability.⁶ We determined how well the citation count of articles that pass basic critical appraisal criteria can be predicted at two years using data available within three weeks of publication, including internal article ratings.⁷

METHODS

The Health Information Research Unit at McMaster University identifies newly published, high quality healthcare evidence (studies and systematic reviews)

and brings them to the attention of practising clinicians through print and electronic media. Over 130 journal titles are read by research associates, who apply criteria to ensure methodological rigour of the articles selected.⁸

Articles that meet standards for methods and content are categorised as being important to any of 59 clinical disciplines. Each article is rated for clinical relevance and newsworthiness by practitioners for each pertinent discipline (three raters per discipline) through the McMaster online rating of evidence system.⁷ Clinical relevance is rated using a Likert scale from 1 “not relevant” to 7 “directly and highly relevant.” Newsworthiness is similar: 1 “not of direct clinical interest” to 7 “useful information, most practitioners in my discipline definitely don’t know this.” Articles with minimum ratings of 4 for each scale are sent as email alerts and a subset of important ones are selected for summary in the three evidence based synoptic journals. Articles published from January to June 2005 that had a minimum average clinical rating of 4 for each scale for at least one clinical discipline were studied. We collected 17 article specific and three journal specific variables for each article (see bmj.com).

Citation counts were collected from the Institute of Scientific Information web of science between 17 April and 8 June 2007, giving almost 24 months of citation accrual (see bmj.com for exclusions).

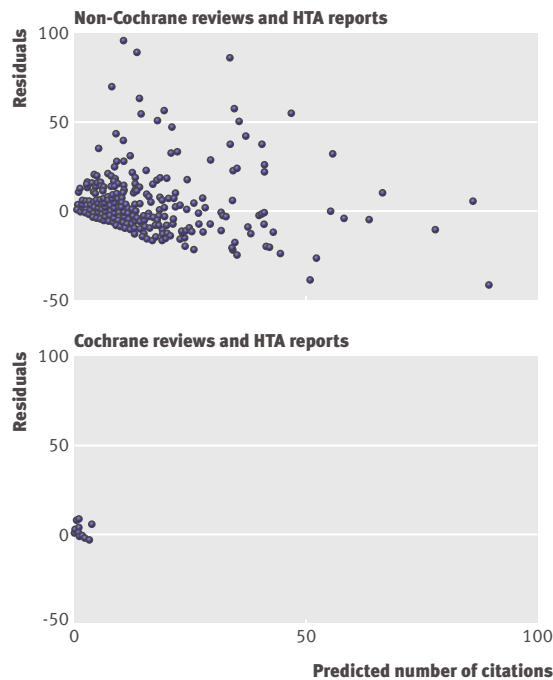
Analysis

Multiple regression used a random split of data into 60:40 derivation-validation datasets and 20 outcome variables. After the regression we tested colinearity with the variance inflation factor whereby variables were removed if the variance inflation value was over 5.⁹ We also tested for outliers using added value plots and for the normality of the regression residuals using residual plots.

To determine the potential of our model for fitting other samples we did a cross validation.^{10,11} Using the regression model from the derivation dataset we predicted values for each case in the validation dataset and computed R^2 values. Subtracting the R^2 generated by applying the derivation model to the validation dataset from the derivation R^2 value provided the “shrinkage on cross validation” of the analysis—measuring how well the derived regression equation matched the actual data from the validation dataset.^{10,11} We used the derivation dataset to determine the sensitivity and specificity of our model in detecting

This article is an abridged version
of a paper that was published on
bmj.com on 21 February 2008.

Cite this article as: *BMJ*
21 February 2008, doi: 10.1136/
bmj.39482.526713.BE



Plot of residuals versus predicted citations for 504 articles from validation dataset (outliers excluded). Perfect agreement is indicated by observations along zero line on vertical axis. Top graph represents 396 articles in validation dataset, excluding 108 articles that are Cochrane reviews or reports from the Health Technology Assessment (HTA) database (bottom graph)

the articles with citations greater than the median and those that were in the top third of cited articles.

The added variable plots for each variable in the original regression indicated several outliers. These plots show cases that exert disproportionate influence on the regression model.⁹ These outliers were articles with extremely high citation counts. Residual plots indicated some non-normality in the model. Because citation counts were highly negatively skewed we transformed them with square roots of the counts. The variance inflation factor indicated that the proportion of articles passing reading criteria per journal in 2005 correlated with the proportion abstracted, and we removed the proportion passing from the analyses.

RESULTS

About 15 000 articles were reviewed from January to June 2005. Of those, 1310 articles passed criteria for methods and had average ratings of at least 4 for clinical relevance and newsworthiness. Forty nine articles were excluded (13 had citation counts >150 and were removed from the regression only) leaving 1261 articles from 105 journals (see *bmj.com* for flow of study data).

Citation counts for the derivation dataset (n=757) varied from 0 to 128 (mean 13.1, median 6) with substantial negative skew; 166 articles had no citations. Cochrane reviews and Health Technology Assessment (HTA) reports were 24% of the sample (n=182).

The resulting multiple regression on the derivation dataset was highly significant (see *bmj.com*): $R^2=0.60$ (95% confidence interval 0.538 to 0.629), $P<0.001$. Nine article specific variables were statistically significant: the number of authors, abstraction in a synoptic journal, clinical relevance score, number of pages, structured abstract, number of cited references, original article, multicentred study, and study about therapy. Two journal specific variables—the number of bibliographic databases in which the journal was indexed and the proportion of articles that were abstracted in 2005—also predicted higher citations.

The hypothesis was that number of pages and a structured abstract would have a positive influence and that being an original study would have a negative influence (see *bmj.com*). These variables had statistically significant counterintuitive results possibly driven by the number of Cochrane reviews (23% in derivation dataset, 20% in validation dataset), which tend to be longer than journal articles and to have structured abstracts. They also had low citation rates (mean 0.46, range 0-9).

To test how well the model fitted another dataset, the derivation regression model created predicted values for each case in the validation dataset. For this validation dataset (n=504) citation counts varied from 0 to 120 (mean 12.9, median 7) with substantial negative skew; 109 articles had no citations. Cochrane reviews and HTA reports accounted for 21% of the sample (n=108).

The R^2 based on the validation dataset was 0.56 (95% confidence interval 0.48 to 0.60, $P<0.001$). The shrinkage on cross validation was low at 0.04, or 4%, suggesting that the model can be generalised to similar datasets.¹¹ A plot of the residuals shows the distribution of the difference in observed and predicted citation counts (figure): average difference between observed and expected counts 4.95 (95% confidence interval 2.03 to 7.86).

Using the validation results (six outliers included) the model had a sensitivity of 83.3% and a specificity of 71.5% for predicting the top half (>7 citations) of cited articles and a sensitivity of 66.1% and a specificity of 82.2% for predicting the top third (>12 citations). The area under the receiver operating characteristic curve was 0.76 (95% confidence interval 0.722 to 0.80) for a threshold set at seven articles—the median (see *bmj.com*). The model did fairly well at discriminating the top performing articles.

DISCUSSION

The number of citations an article accrues within two years can be predicted with about 60% surety using data within three weeks of publication. This ability to predict citation counts is high. One study was only able to predict 14% of the variance for the annual citation rates using papers on emergency medicine and 12 possible predictor variables,¹² whereas another study predicted 20% using articles from *JAMA*, the *Lancet*, and the *New England Journal of Medicine*.¹³ Our sample,

however, represents a select group of articles that passed methodological criteria and the greater predictability could result from lower variability in the dataset.

We have also shown that physician rated clinical relevance at the time of publication is related to citation counts at two years.

Quality of studies has been shown to be weakly or moderately related to citation counts.² We could not assess the influence of quality of research methods because our articles passed basic methodological criteria, and we had not graded their quality.

Cochrane reviews and HTA reports do not have impact factors. Given the importance of such articles we chose to include them and not to use the impact factor as a predictor variable. These systematic reviews of evidence are typically lengthy and have structured abstracts. Their average citation rates were low compared with the other study articles. We believe that this underlies our findings of negative correlations between citation counts and number of pages, structured abstract, and review articles. Without Cochrane reviews and HTA reports in our analysis, comparisons between the number of pages and original article versus review articles were no longer statistically significant, although structured abstracts remained negatively correlated with citation counts. As we had no prior hypothesis on the influence of Cochrane reviews and HTA reports on citation counts we included them. We are aware that the reduced variability among these publications improved the performance of our regression.

The positive association between clinical ratings and citation counts could be because early dissemination actually leads to higher citation rates, not merely predicts citation. We could not assess this possibility, but the intended targets of our dissemination process are practising clinicians, not scientific authors. Thus our findings support an association that constitutes “criterion validity,” in that ratings predict an accepted measure of research merit.

The strengths of this study include the large number of articles (n=1261), the magnitude of the association in the derivation dataset, and the agreement between the results from the derivation and validation datasets. We have shown a statistically significant relation between ratings of the clinical relevance of an article and its citation count.

Several weaknesses are present. We included only 105 of the most important clinical journals, therefore our results may not be readily transferable to articles in less important clinical journals or basic science articles or journals. Also, we included only those clinical articles passing criteria for critical appraisal and they represent a small proportion of articles published in any journal. Such a select sample would have reduced variability resulting in greater predictability. Including the Cochrane reviews and HTA reports also reduced the variability and led to greater predictability.

We used 20 journal specific and article specific characteristics of 1261 articles from 105 clinical journals to predict citation counts at two years. Eleven remained significant in our regression model.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Citation counts are markers of an article's importance but are not available for months after publication

Research shows that various attributes of an article are related to higher citation rates, but the predictive value of these factors is limited

WHAT THIS STUDY ADDS

Features of methodologically sound articles predicted citation counts with higher reliability than previously found

Ratings of clinical relevance by practising clinicians are significantly associated with citation counts at two years

Therefore we can predict citation counts of methodologically sound clinical studies and review articles at two years with surety using data available early after publication.

Contributors: See bmj.com.

Funding: No external funding.

Competing interests: The authors are employees of McMaster University. McMaster University owns the intellectual property for some of the processes described in the study including the McMaster premium literature service and the McMaster online rating of evidence (MORE), which are used to appraise and select articles for *ACP Journal Club*, *Evidence-Based Medicine*, *Evidence-Based Nursing*, *bmjupdates+*, *BMJ Clinical Evidence*, Medscape Best Evidence Alerts, Physicians Information and Education Resource, and Harrison's Practice. All of the coauthors are or have been employed in part by contracts between McMaster University and the publishers of these information services.

Three BMJ Group products are mentioned in this paper: *Evidence-Based Medicine*, *Evidence-Based Nursing*, and *bmjupdates+*.

Ethical approval: Not required.

Provenance and peer review: Not commissioned; externally peer reviewed.

- Adams AB, Simonson D. Publication, citations, and impact factors of leading investigators in critical care medicine. *Respir Care* 2004;49:276-81.
- West R, McLlwaive A. What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality. [see comment]. *Addiction* 2002;87:501-4.
- Campbell FM. National bias: a comparison of citation practices by health professionals. *Bull Med Libr Assoc* 1990;78:376-82.
- Bormann L, Hans-Dieter D. Multiple publication on a single research study: does it pay? The influence of number of research articles on total citation counts in biomedicine. *J Am Soc Inf Sci* 2007;58:1100-7.
- Figg WD, Dunn L, Liewehr DJ, Steinberg SM, Thurman PW, Barrett JC, et al. Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy* 2006;26:759-67.
- Eysenbach G. Citation advantage of open access articles. *PLoS Biol* 2006;4:e157.
- Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski NL, Jedraszewski D, et al. A second order of peer review: a system to provide peer review of the medical literature for clinical practitioners. *JAMA* 2006;296:1801-8.
- Wilczynski NL, McKibbin KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *Medinfo* 2001;10(Pt 1):390-3.
- Hamilton LC. *Statistic with STATA, updated for version 9*. Belmont, CA: Thomson Books/Cole, 2006.
- Kleinbaum DGK, Kupper LL, Muller KE. *Applied regression analysis and other multivariable methods*. 2nd ed. Boston: PWS-Kent Publishing, 1988.
- Osborne JW. Prediction in multiple regression. *Pract Assess Res Eval* 2000;7(2).
- Callahan M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA* 2002;287:2847-50.
- Kulkarni AV, Busse JW, Shams I. Characteristics associated with citation rate of the medical literature. *PLoS ONE* 2007;2:e403.

Accepted: 28 January 2008