


ORIGINAL ARTICLE

Predicting future citation counts of scientific manuscripts submitted for publication: a cohort study in transplantology

Michael Kossmeier¹ & Georg Heinze² 

1 Department of Basic Psychological Research and Research Methods, School of Psychology, University of Vienna, Vienna, Austria

2 Section for Clinical Biometrics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria

Correspondence

Georg Heinze PhD, Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
Tel.: +4314040066880;
fax: +4314040066870;
e-mail:
georg.heinze@meduniwien.ac.at

SUMMARY

Citations are widely used for measuring scientific impact. The goal of the present study was to predict citation counts of manuscripts submitted to *Transplant International* (TI) in the two calendar years following publication. We considered a comprehensive set of 21 manuscript, author, and peer-review-related predictor variables available early in the peer-review process. We also evaluated how successfully the peer-review process at TI identified and accepted the most promising manuscripts for publication. A developed predictive model with nine selected variables showed acceptable test performance to identify often cited articles (AUROC = 0.685). Particularly important predictors were the number of pages, month of publication, publication type (review versus other), and study on humans (yes versus no). Accepted manuscripts at TI were cited more often than rejected but elsewhere published manuscripts (median 4 vs. 2 citations). The predictive model did not outperform the actual editorial decision. Both findings suggest that the peer-review process at TI, in its current form, was successful in selecting submitted manuscripts with a high scientific impact in the future. Predictive models might have the potential to support the review process when decisions are made under great uncertainty.

Transplant International 2019; 32: 6–15

Key words

citation prediction model, citations, editorial decision, journal impact factor, journal peer review, predictive validity, review evaluation

Received: 6 April 2018; Revision requested: 24 April 2018; Accepted: 10 June 2018;

Published online: 22 July 2018

Introduction

The number of citations a scientific publication receives is generally seen as a measure of scientific impact and importance [1]. Therefore, journal editors are interested in identifying, amongst the many submissions they receive, manuscripts with the potential to receive many citations. In particular, they aim to increase the journal's prestige, often expressed in a single measure, the *Journal Impact Factor* (JIF) [2,3].

In the past, citations to scientific articles were used to evaluate the review process of scientific journals [4]. Generally, many more manuscripts are submitted to scientific journals for publication than journals are able to publish. The peer-review process is a mechanism to decide which submitted manuscripts are worth publishing and which should be rejected. Naturally, the question arises how successfully journals and their peer-review processes – on average – identify the best manuscripts to publish. One proven evaluation approach is

to compare citation counts of manuscripts accepted and published by a journal with citations to rejected manuscripts that were later published elsewhere [5–8].

We developed a model to predict future citations of manuscripts submitted to *Transplant International* (TI) for this study. Although various potentially prognostic factors for citations were considered in prior research, early indicators for the citation success of manuscripts have hardly ever been used, particularly those variables which arise from the journal peer-review process [9]. In order to fill this gap, we utilized TI review data available to the editorial board at the time of their first decision to reject or further consider a manuscript. We also critically evaluated the journal peer-review process of TI by comparing citation counts of manuscripts which were published by TI with citation counts of submitted manuscripts which were rejected but later published elsewhere. Finally, we compared the actual performance of the editorial board with the hypothetical performance of our predictive model.

Methods

Methodological decisions and strategies regarding data acquisition, data management, and data analysis were specified in a study protocol. If not explicitly stated otherwise, all methodological decisions were made prior to data acquisition and are briefly described in the following sections.

Transplant International uses *ScholarOne*, an online application tool to manage the submission of manuscripts and the peer-review process. We used this system to retrieve data from the review process for our study. Complementary information was retrieved from other sources (see below). We included original studies or reviews and excluded letters to the editor, invited commentaries, and conference abstracts because they arguably represent special cases in the journal (peer-) review process. Furthermore, we excluded case studies from analysis because they are no longer published in TI.

Study period and study cohorts

Two distinct ‘cohorts’ of manuscripts submitted to TI were used (Table 1). A training cohort was used to build the predictive model and a test cohort exclusively served to externally validate the model and evaluate the review process. The training cohort consisted of all 259 submitted manuscripts of all eligible articles that were published in TI in the years 2011 or 2013. For validation, we composed a cohort consisting of manuscripts published in TI and manuscripts rejected by TI which later were published elsewhere. First, we randomly selected 75 manuscripts which were published by TI in 2012. Second, a random sample of 200 rejected manuscripts at TI was followed up by searching in Web of Science for possible later publication in another journal. Of these 200 rejected manuscripts, 68 eligible manuscripts could be tracked down as published papers in other journals and were added to the test cohort. Details on search strategies, eligibility criteria, and the selection of the test cohort are described in the Appendix S1.

Dependent variable: Impact factor relevant citations

The rate at which published articles receive citations after they were published is generally not constant over time [10]. Thus, for every article, we extracted the number of citations received in the 2 years following its publication year from Web of Science, which are exactly the citations relevant for the computation of the JIF (further details can be found in the Appendix S1).

Predictor variables

Overall, we considered 22 distinct covariates as relevant to predicting the future impact factor relevant citations of a manuscript. One variable had to be excluded from our data analysis because of poor data quality (corresponding author h-index). Of the remaining 21 predictors (Table 2), 17 were regularly available at the time of the first decision in the review process. Four variables (number of pages, publication month, industry funding

Table 1. Manuscript samples used for analysis.

Training cohort (N = 259)	Test cohort (N = 143)	
	Accepted (N = 75)	Rejected (N = 68)
All 2011 and 2013 published eligible articles in TI	Random sample of all 2012 published eligible articles in TI	Random sample of 200 rejected manuscripts out of which 68 were later published elsewhere

Table 2. Considered covariates to predict future citation counts.

Predictor	Data source ^a	Training cohort (N = 259)	Test cohort – accepted (N = 75)	Test cohort – rejected (N = 68)
Manuscript information				
Study on humans (yes versus no) ^b	A	Yes: 221 (85%) No: 38 (15%)	Yes: 68 (91%) No: 7 (9%)	Yes: 59 (88%) No: 8 (12%)
Randomized study (yes versus no) ^b	A	Yes: 44 (17%) No: 215 (83%)	Yes: 10 (13%) No: 65 (87%)	Yes: 10 (15%) No: 58 (85%)
Publication type (review versus other)	A	Review: 39 (15%) Other: 220 (85%)	Review: 9 (12%) Other: 66 (88%)	Review: 3 (4%) Other: 65 (96%)
Meta-analysis (yes versus no)	A	Yes: 4 (2%) No: 255 (98%)	Yes: 1 (1%) No: 74 (99%)	Yes: 1 (1%) No: 67 (99%)
Organ focus (kidney versus liver versus heart versus lung versus other/none)	A	Kidney: 82 (32%) Liver: 78 (30%) Heart: 14 (5%) Lung: 11 (4%) o/n: 74 (29%)	Kidney: 34 (45%) Liver: 22 (29%) Heart: 2 (3%) Lung: 4 (5%) o/n: 13 (17%)	Kidney: 30 (44%) Liver: 23 (34%) Heart: 5 (7%) Lung: 2 (3%) o/n: 8 (12%)
Industry funding (yes versus no)	WoS	Yes: 34 (13%) No: 225 (87%)	Yes: 14 (19%) No: 61 (81%)	Yes: 11 (16%) No: 57 (84%)
Manuscript title word count	A	14 (11, 16.5)	14 (10.5, 17)	15 (13, 17)
Sample size (if applicable) ^c	A	160.5 (55.25, 459)	110 (50, 342)	114 (59, 274)
Number of cited references	WoS	33 (26, 46)	34 (25.5, 45)	28 (18.75, 38)
Number of pages	WoS	9 (8, 10)	9 (7, 10)	8 (6, 9)
Article month of publication	WoS	6 (3, 9)	7 (3, 9)	7 (5.75, 10)
Author information				
Corresponding author institution ranking ^d	SIR	160 (139.5, 187.5)	160 (133, 177.5)	150 (95, 180)
Number of authors	R	8 (5, 10)	7 (5, 10)	7 (5, 9)
Number of different (unique) author institutions	R	2 (1, 3)	2 (1, 3)	1 (1, 2)
Review process information				
Excess reviewers invited	R	3 (2, 5)	4 (2, 5)	3 (2, 5)
Proportion of reviews completed	R	0.43 (0.3, 0.6)	0.5 (0.33, 0.67)	0.6 (0.5, 0.67)
Mean number of days for review completion	R	12.25 (9.84, 15)	12.33 (9.71, 14.71)	12 (10.19, 13.50)
Days between submission and first decision	R	35 (27, 46)	32 (25.5, 40)	33.5 (28, 40.25)
Originality score ^e	R	3 (2.55, 3.25)	3 (2.5, 3)	2 (2, 3)
Scientific quality score ^e	R	2.75 (2.5, 3)	3 (2.5, 3)	2 (2, 2.5)
Presentation score ^e	R	3 (2.5, 3)	3 (2.5, 3)	2.5 (2, 3)

Univariate descriptive statistics are number (percentage) of categories for categorical variables and median (first quartile, third quartile) for continuous variables.

^aData source: A = manuscript abstract, R = review database, WoS = Web of Science, SIR = SCImago Institutions Rankings (<http://www.scimagoir.com>).

^bBecause the majority of randomized studies in the sample were experimental animal studies, we additionally considered an interaction of the two dichotomous variables Study on humans and Randomized study.

^cThe sample size was only considered for original studies on human samples. Because sample size was severely right skewed we transformed by log-base-10 and centred by the mean of the training data. All manuscripts without (eligible) sample sizes (animal studies, reviews) were set to zero (i.e. to the mean of the training cohort).

^dThe SCImago Institutions Rankings (SIR, <http://www.scimagoir.com>; accessed 8 March 2018) World Report contains different yearly published global rankings and scores of over 2000 research institutions worldwide. We used the *Normalized Impact* of the corresponding author's institution in the latest completed year prior to the first decision (times 100). A *Normalized Impact* score of 100 corresponds to the world average of all citations to published research originating from an institution. A score of 80 means that published research originating from an institution is cited 20% below world average. A score of 130 means published research from an institution is cited 30% above world average.

^eMean score of reviewer ratings. Every reviewer rates the submitted manuscript on three dimensions (Originality, scientific quality and presentation), coded from 1 (Poor) to 4 (Excellent).

and number of references) were extracted from the published articles rather than the submitted manuscripts, because they are either in high agreement with or are already determinable based on the information in a manuscript. A detailed description of all covariates can be found in the Appendix S1.

Data analysis

For the estimation of the model to predict future citations and to select variables we exclusively used the training cohort ($N = 259$), whilst the test cohort ($N = 143$) remained strictly reserved for model validation. All analyses were conducted using the statistical software R and extensions [11–15].

Predictive modelling approach

We used negative binomial regression to model the number of citations. First, we fitted a full model with all covariates to obtain effect estimates and 95% confidence intervals (Appendix S2). Second, we used LASSO negative binomial regression to select covariates relevant for prediction and to increase predictive accuracy via shrinkage [16,17]. The optimal amount of penalization was determined by maximizing the 10-fold cross-validated likelihood statistic. Predictors were centred and standardized for LASSO estimation. Model stability was assessed by bootstrap inclusion fractions (BIF), i.e., evaluating how often each variable was selected when repeating the LASSO selection in 1000 bootstrap resamples (for details and further sensitivity analyses see Appendix S3).

Model validation and validation of editorial as well as model-based acceptance decisions

The agreement of predicted and observed citations was assessed using calibration plots. Explained variation was estimated with the correlation coefficient of predicted and observed values, a universally applicable and interpretable measure [18]. TI had a JIF 2015 (published 2016) of 2.835. Therefore, manuscripts could be classified as positively (negatively) contributing to the impact factor if their citation counts received in the 2 years after the publication year exceeded (fell below) 5.670 (= impact factor \times 2). We computed the area under the receiver operating characteristic curve (AUROC) to assess how well the model correctly discriminates between often cited and rarely (or not) cited manuscripts. Sensitivity and specificity were estimated by the

relative frequencies with which often cited and rarely cited manuscripts were correctly identified, respectively. Bootstrapping was used to perform internal validation for optimism correction [19] (see Appendix S3 for further details). The test cohort was used for external validation.

To transform the model predictions of citation counts into a recommendation to accept or reject, we chose 5.670 (= impact factor \times 2) predicted citations as a natural cut-off value.

The test cohort consisted of accepted manuscripts as well as of 'rejected-but-published-elsewhere' ones. This allowed for computation of the ability of the editorial decision to accept often cited articles (sensitivity) and to reject rarely cited manuscripts (specificity), as well as to compare this with the model recommendation. Because of the specific design of the test cohort consisting of pre-defined proportions of rejected (48%) and accepted (52%) papers, sensitivity and specificity estimates must be interpreted with these proportions in mind.

Results

Descriptive statistics of all considered predictor variables for the training and test cohorts can be found in Table 2. Figure 1 shows the distributions of *impact factor relevant citations* for the training cohort (Fig. 1a) and the test cohorts (Fig. 1b,c). In the training cohort, the average (median) citation count was 6.3 (5). The maximum number of citations achieved by one manuscript was 47, and 15 (5.8%) papers were not cited in the 2 years after the publication year. In the test cohort (accepted and rejected combined), the average (median) citation count was 4.2 (3).

Predicting future citation counts

LASSO selected nine predictors for future citation counts: publication type (review versus other), number of pages, month of publication, study on humans (yes versus no), days between manuscript submission and first decision, mean number of days for review completion, number of different (unique) author institutions, number of authors and number of cited references. Thus, five predictors were directly related to information from the manuscript, two described authors and two the review process. The regression coefficients, resulting count multipliers, standardized regression coefficients (β) and BIF are given in Table 3. The most important predictors were all manuscript-related:

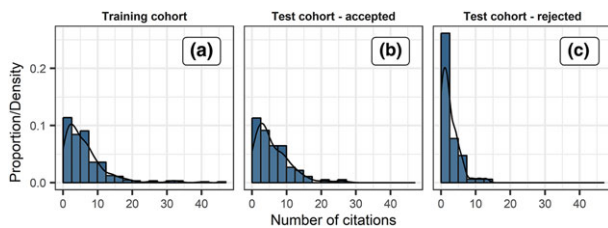


Figure 1 Histogram and kernel density estimator of citations in the 2 years after the publication year for manuscripts in the training cohort (a), and test cohort (b,c).

Type = Review ($\beta = 0.08$, BIF = 0.957), number of pages ($\beta = 0.088$, BIF = 0.935), month of publication ($\beta = -0.052$, BIF = 0.901) and study on humans ($\beta = 0.069$, BIF = 0.802). BIF were indeed highest for the nine actually selected variables. All nine selected variables had inclusion fractions over 60% (Table 3). Therefore, the set of selected variables seemed to be quite robust in terms of random sampling variability.

Internal validation of the LASSO negative binomial regression model using the training cohort and optimism corrected performance measures showed a correlation of observed and fitted values of $r = 0.205$. The ability to discriminate between often cited and rarely cited manuscripts was estimated as AUROC = 0.636.

In external validation using the test cohort, fitted and observed values were even more highly correlated ($r = 0.422$). The model showed a moderate ability to discriminate between often cited and rarely cited manuscripts within the test cohort (AUROC = 0.685, Fig. 2a). Because of the strong shrinkage of regression

coefficients compared with the full, unpenalized model (Appendix S2), the predicted values were all close to the mean citation count and their variance was considerably smaller than that of the observed ones (Fig. 2b).

Evaluation of the editorial decision

In the test cohort, manuscripts that were published in TI received, on average, more impact factor relevant citations (mean 5.84, median 4) than manuscripts that were rejected and later published elsewhere (mean 2.52, median 2; Fig. 3a).

Of all 37 often cited manuscripts in the test cohort, 29 (78.4%) were accepted by TI, suggesting a sensitivity of the editorial decision of 0.784. 106 manuscripts were rarely or not cited, and of these 60 (56.6%) were rejected and 46 were accepted by TI corresponding to a specificity of 0.566. The positive and negative predictive values of acceptance and rejection were 0.387 and 0.883, respectively. In other words, of all accepted manuscripts 38.7% (29 of 75) turned out to be often cited, whereas of all the rejected manuscripts 88.3% (60 of 68 manuscripts) performed poorly. The odds of a later often cited manuscript to be accepted by TI was nearly five times greater than for a manuscript that was rarely cited [odds ratio 4.728, 95% CI (1.977, 11.307)].

As a new methodological approach to evaluate the review process, we compared the ability of the editorial board to identify and accept later often cited articles with decisions solely based on our prognostic model. We hypothetically assumed that the model rejects

Table 3. Coefficients of the LASSO negative binomial regression model to predict the number of impact factor relevant citations.

Variable (x_j)	Coefficient b_j	Count multiplier e^{b_j}	Standardized coefficients β_j	Bootstrap inclusion fraction (rank)
Intercept b_0	1.262	3.531	1.797	1 (n.a.)
Type = Review ($x_j = 1$), other ($x_j = 0$)	0.223	1.250	0.080	0.957 (1)
Pages	0.037	1.038	0.088	0.935 (2)
Month (1–12)	−0.015	0.985	−0.052	0.901 (3)
Human = Yes ($x_j = 1$), other ($x_j = 0$)	0.196	1.216	0.069	0.802 (4)
Days to first decision	−0.001	0.999	−0.008	0.685 (5)
Number of author institutions	0.016	1.016	0.050	0.676 (6)
Mean days to review completion	−0.003	0.997	−0.013	0.665 (7)
Number of authors	0.003	1.003	0.012	0.661 (8)
Number of references	0.002	1.002	0.052	0.607 (9)

Citations can be predicted by $e^{(b_0 + \sum b_j x_j)}$, where x_j is the value of the predictor variable. Count multipliers e^{b_j} can be interpreted as multiplicative effect on the estimated citation count for a one unit increase in the respective variable (e.g., holding all other variables constant, the estimated citation count is 25% higher for reviews compared with other studies). The inclusion fraction is the proportion of 1000 bootstrap samples where LASSO models included that coefficient (see Appendix S3 for details).

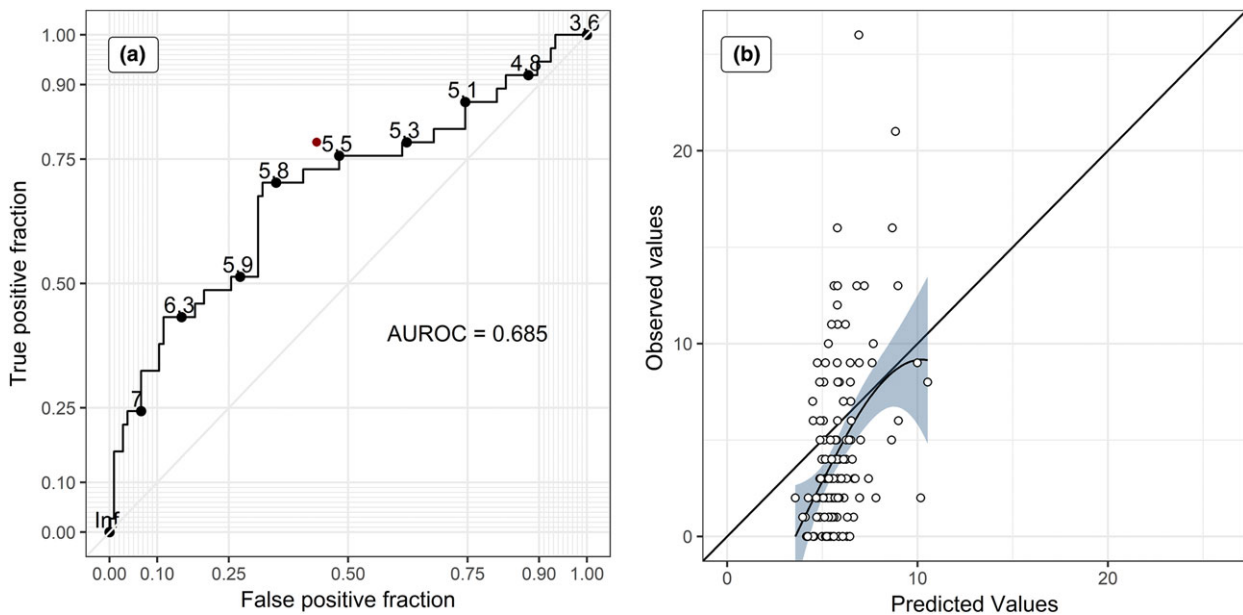


Figure 2 ROC-curve and calibration plot for predicting the test cohort (external validation). (a) The ROC curve of the LASSO model for the test cohort indicating the ability of the model to discriminate between often cited and rarely cited manuscripts for different cut-off values of predicted citations. The red dot indicates the specificity and sensitivity of the editorial decision within the test cohort. (b) Scatter plot of predicted citations by the model and actually observed citations in the test cohort. For a good calibration (no systematic over- or underestimation) the points should lie on or close to the 45° line. To help the visual assessment, a non-parametric loess regression with (pointwise) 95% confidence band is shown.

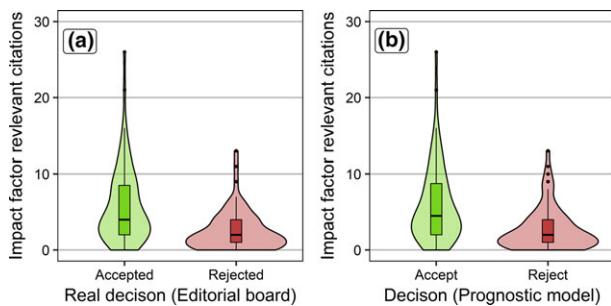


Figure 3 Impact factor relevant citation distribution of manuscripts in the test cohort. (a) Distribution of impact factor relevant citations for submitted manuscripts which were accepted by *Transplant International* for publication compared to rejected but published elsewhere manuscripts. (b) Impact factor relevant citations of (hypothetically) accepted and rejected manuscripts if the acceptance decision is based on the predictions of the prognostic model.

manuscripts with predicted values smaller than or equal to our cut-off (5.670, twice the impact factor 2015), whilst it accepts predicted values above the cut-off. In the test cohort this decision rule leads to 77 (53.8%) rejections and 66 (46.2%) acceptances. The mean (median) actual citation count of these papers was 2.779 (median 2) and 5.985 (median 4.5), respectively (Fig. 3b).

The sensitivity of the model was 0.703 (26 of 37 often cited manuscripts were accepted) whilst the specificity was 0.623 (66 of 106 rarely cited manuscripts were rejected).

The positive and negative predictive values of acceptance and rejection decisions were 0.394 and 0.857, respectively. The odds ratio for acceptance versus rejection with respect to high citation count was 3.9 [95% CI (1.74, 8.74)]. Acceptance decisions based on the model differed from the actual editorial decision in 34.3% of all test cohort manuscripts. Remarkably, decisions based on the model performed almost as well as the actual editorial decisions. Performance numbers are summarized in Table 4.

Discussion

In a recent review 198 papers published between 2000 and early 2015 could be identified which examined which variables might predict and explain citations to scientific articles [9]. The identified studies varied by scientific discipline, by the set of explanatory variables considered and by their methodology. Overall, 28 paper related (e.g., paper length), author related (e.g., number of authors) and journal related factors (e.g., journal scope) were considered [9].

In medical research, the number of cited references [20,21], the number of authors [21,22], and the JIF of the journal the article was published in [20,23] were consistently important. In addition, reviews were on average more often cited than original research articles [21] and funding from pharmaceutical industry, especially when

Table 4. Performance numbers for the ability to discriminate between often cited and rarely or not cited submitted manuscripts in the test cohort.

Test cohort	Positive predictive value	Negative predictive value	Sensitivity	Specificity	Odds ratio
Editorial decision	0.387	0.883	0.784	0.566	4.728
Model-based decision	0.394	0.857	0.703	0.623	3.900

results were in favour of industry, was shown to be associated with higher citation counts [24,25].

However, for many other possibly explanatory variables, findings were rather inconsistent and seemed to depend on the studies context (for an overview see [26]).

Despite the large number of studies and number of explanatory variables investigated only two studies considered variables relating to the journal peer review. Articles with longer review times until acceptance often had higher citation counts [27], whilst reviewers' assessment of the scientific quality of articles was not found to be significantly correlated with citation counts [28].

We considered a comprehensive set of 21 variables related to the article, authors, and the peer-review process to predict citation counts in the 2 years following the publication year.

For manuscripts submitted to TI, a higher number of article pages predicted higher citation counts. Remarkably, in a similar study from medical literature, the opposite was observed with shorter papers receiving more citations [21]. However, in that study, articles from 105 different scientific journals were used. Often there is a large variance between journals in the number of pages of their published articles. Thus, the number of pages might carry information about the journal in which the article was published (e.g., the journal prestige).

For model building, we exclusively used articles published in TI. Hence, we can exclude confounding with journal prestige.

Several reasons for the association of the number of pages and future citations are possible. First and most trivially, lengthier articles might contain more content they can be cited for. Second, journal space is limited and therefore promising research papers might be assigned larger portions of this space. Third, TI charges a fee for every page exceeding seven pages for original articles. Therefore, papers with more pages might be an indicator for higher resources of the author or the author's institution.

It is well known that review articles are cited more often, on average, than original articles and we also observed this in our study [29]. Reviews are often of broader interest and can therefore attract citations by a more diverse follow up literature than original studies.

Furthermore, it is often more efficient to cite one review summarizing a large number of original research articles than citing a large number of original research articles themselves. Therefore, our results agree with the widespread belief that – measured in citations – reviews generally have a higher scientific impact. However, there are also a large number of published reviews that are not or are hardly ever cited [30] which would suggest that quality might also play an important role here.

Finally, the month of publication of articles in TI was related to the number of citations in the two calendar years following publication. This association may vanish if considering a longer time frame, but is highly relevant for a journal's JIF. Citation rates of individual articles often increase only after some delay, reflecting the time needed by the citing articles to get published [10,31]. Thus, articles published early in their publication year have a higher chance to overcome the mentioned citation delay and contribute more citations to the JIF. Because articles are only considered for the JIF after their publication in print this lead to the practice to electronically publish accepted articles far ahead of print [32] as well as to publish promising manuscripts early in the year [33] to boost the JIF. Although the month of publication is not yet fixed at the time of first decision (at which the prediction model could be used), it can be roughly estimated from the usual times from first decision to publication. Our model may also be used to evaluate the impact of publishing in a certain month, e.g. December or January of the subsequent year.

A higher number of authors or a higher number of unique author institutions leads to a higher predicted citation count, in agreement with the literature [22]. Many authors from different institutions might, on average, lead to higher quality and highly visible research, with multi-institutional studies indicating resource-intensive research. A higher number of authors might also increase the likelihood of future self-citations [34]. In addition, we found that studies on humans were predicted to attract more citations than studies on animals. In contrast to findings in earlier studies which investigated univariate associations, only a weak effect of the number of references was observed in our

multivariable model. This could be explained by confounding with the effects of the number of pages and the publication type (review or other).

Manuscripts with shorter mean review duration and mean days from submission to a first decision were predicted to have more citations. Short review and decision times might be an indicator of quality. However, we found no association between quality ratings and review time with citation counts (data not shown). Review and decision times might also be an indicator of the complexity or degree of specialization of a manuscript, with manuscripts on complex or niche topics probably taking on average longer to review but attracting fewer citations. Other authors found that manuscripts with longer total review times attract more citations [27]. Our findings are not necessarily inconsistent with this because instead of total time in review we considered the time from the submission to a first decision in the review process.

Remarkably, quality, originality and presentation scores of manuscripts rated by peer reviewers were not strongly associated with future citation counts. This can either mean that manuscripts generally improve between initial submission and publication, with suggestions to improve the manuscript made by peer reviewers being one driving factor. Alternatively, information on these scores may already be covered by other predictors in the model (e.g., number of pages) or may not be relevant once a particular quality threshold has been reached. This is in accordance with previous findings [28].

Articles published by TI achieved considerably higher mean and median citation counts than rejected ones but elsewhere published manuscripts. The editorial board identified most of the later often cited manuscripts of the test cohort, but performed slightly worse in identifying and rejecting rarely cited manuscripts. Still, the odds of accepted manuscripts to be often cited was found to be nearly five times higher than the corresponding odds for rejected but elsewhere published manuscripts.

Beyond that, we proposed and demonstrated a new evaluation approach by comparing the performance of our predictive model with the editorial decision at TI. The hypothetical decisions based on the predictive model did not outperform the actual editorial decision. This further supports the conclusion that the prognostic validity of the peer review process at TI is rather high.

Retrospective predictive research is observational by nature and this study is no exception. Our study might give hints on possible factors favouring high citations and might help to identify indicators for high citation counts in the future but supplies only little evidential value with respect to true causal pathways.

Citation counts are only one of several ways to judge research performance and most certainly only one aspect of research impact. For instance, high quality replication studies are essential for the accumulation of strong empirical evidence in medicine and other empirical sciences but such studies might on average be cited less often than studies with novel findings.

We considered citations to articles in the 2 years after their publication year in line with the computation of the JIF. This is a rather short timeframe and can be seen as a measure of short-term impact. Articles that are not cited heavily until several years after their publication (so called *sleeping beauties* [35]) cannot be identified with this approach. However, popular metrics, such as the 2-year JIF, exclusively depends on early citations.

A further limitation is the rather small number of manuscripts used for this study. Because the focus was on articles published in one specific journal, it was not feasible to arbitrarily increase the sample size in order to warrant contemporary conclusions.

In addition, rejected but elsewhere published manuscripts are a selected sample of all rejected manuscripts. Compared with other studies the relative frequency of successfully retrieving such manuscripts was rather low, potentially increasing the problem. For instance, in their pioneering study, Bornmann and Daniel report that over 90% of rejected manuscripts at *Angewandte Chemie International Edition*, were found to be published elsewhere 6 years later, and 75% of these rejected but elsewhere published manuscripts differed only marginally in their content [5]. For a random sample of submitted but rejected manuscripts at TI in the timeframe from December 2011 to July 2013, we found 68% to be published elsewhere as of June 2016. Only 34% were published by 2013 and included in our analysis. We used this relatively strict inclusion rule to increase the likelihood that articles published elsewhere corresponded to a high degree to the manuscript submitted to TI.

Finally, we could not include a measure for author prestige or prior author productivity in our analysis. We initially planned to include the h-index of the corresponding author in the year before the first decision but had to exclude this variable because of insufficient data quality.

Nevertheless, to the best of our knowledge, this is the first study investigating early indicators for citation success of submitted manuscripts at a journal for publication. For this purpose, we used a most comprehensive set of variables that were related to the manuscript, to authors and to peer-review, and were available in the early peer-review process. Access to some relevant

variables from the review process allowed us to consider several peer-review-related predictors for the first time.

A further strength is that we utilized a modern multi-variable modelling approach to obtain predictions with high accuracy. A large number of prior studies did not utilize this superior approach, or did not consider prediction as the main purpose of model building [9].

We evaluated the peer-review process at the journal TI by comparing the citation impact of accepted and published manuscripts with rejected but elsewhere published manuscripts. Our results confirmed that the review process at TI fulfils its role to identify promising manuscripts. Hence, we contributed to the still understudied subject of prognostic validity of journal peer-review [4] with data from transplantology. In our study, a prediction formula with nine predictor variables was identified and performed similarly as, but not in perfect agreement with the editorial board. Predictive models such as ours might, therefore, have the potential to support decision makers at journals for cases where the decision to ultimately accept or reject a submitted manuscript is made under high uncertainty.

Authorship

Both authors designed the study, collected data, analyzed data, interpreted results, wrote the paper, and revised the paper.

Funding

The authors have declared no funding.

Conflict of interest

Georg Heinze is statistical editor of *Transplant International* since 2015. He was not involved in any editorial decisions on the manuscripts analysed in this paper.

Acknowledgements

We appreciate the help of Lukas Schalleck and Martin Meyrath from the editorial office of *Transplant International* in preparing parts of the data used in this study. We thank Gretchen Simms for comments on an earlier version of the manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Data acquisition and data handling: details.

Appendix S2. Additional results.

Appendix S3. Sensitivity analyses of our prognostic model.

REFERENCES

- Bornmann L, Daniel HD. What do citation counts measure? A review of studies on citing behavior. *J Doc* 2006; **64**: 45.
- Garfield E. The history and meaning of the journal impact factor. *JAMA* 2006; **295**: 90.
- Reuters T. Journal Citation Reports® 2015. Retrieved September, 2016, from <https://jcr.incites.thomsonreuters.com/>
- Bornmann L. Does the journal peer review select the “best” from the work submitted? The state of empirical research. *IETE Tech Rev* 2010; **27**: 93.
- Bornmann L, Daniel HD. Selecting manuscripts for a high-impact journal through peer review: a citation analysis of communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere. *J Am Soc Inf Sci Technol* 2008; **59**: 1841.
- Bornmann L, Daniel HD. Extent of type I and type II errors in editorial decisions: a case study on *Angewandte Chemie International Edition*. *J Informetr* 2009; **3**: 348.
- Ophhof T, Furstner F, van Geer M, Coronel R. Regrets or no regrets? No regrets! The fate of rejected manuscripts. *Cardiovasc Res* 2000; **45**: 255.
- McDonald RJ, Cloft HJ, Kallmes DF. Fate of manuscripts previously rejected by the *American Journal of Neuroradiology*: a follow-up analysis. *Am J Neuroradiol* 2009; **30**: 253.
- Tahamtan I, Afshar AS, Ahamdzadeh K. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics* 2016; **107**: 1195.
- Wang D, Song C, Barabási AL. Quantifying long-term scientific impact. *Science* 2013; **342**: 127.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2016.
- Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York, NY: Springer, 2002.
- Wang Z, Ma S, Wang CY. Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biom J* 2015; **57**: 867.
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis*, 2nd edn. New York, NY: Springer, 2016.
- Sachs MC. plotROC: generate useful ROC curve charts for print and interactive use. R package, 2016.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodol)* 1996; **58**: 267.
- Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat Neerl* 2001; **55**: 17.
- Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. *Stat Med* 2000; **19**: 1771.
- Harrell FE. *Regression Modeling Strategies*, 2nd edn. New York, NY: Springer, 2015: 103–126.
- Davis PM, Lewenstein BV, Simon DH, Booth JG, Connolly MJ. Open access publishing, article downloads, and

- citations: randomised controlled trial. *BMJ* 2008; **337**: a568.
21. Lokker C, McKibbin KA, McKinlay RJ, Wilczynski NL, Haynes RB. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ* 2008; **336**: 655.
 22. Figg WD, Dunn L, Liewehr DJ, *et al.* Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy* 2006; **26**: 759.
 23. Fu LD, Aliferis CF. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics* 2010; **85**: 257.
 24. Kulkarni AV, Busse JW, Shams I. Characteristics associated with citation rate of the medical literature. *PLoS One* 2007; **2**: e403.
 25. Farshad M, Sidler C, Gerber C. Association of scientific and nonscientific factors to citation rates of articles of renowned orthopedic journals. *Eur Orthop Traumatol* 2013; **4**: 125.
 26. Onodera N, Yoshikane F. Factors affecting citation rates of research articles. *J Assoc Inf Sci Technol* 2015; **66**: 739.
 27. Hilmer CE, Lusk JL. Determinants of citations to the agricultural and applied economics association journals. *Rev Agric Econ* 2009; **31**: 677.
 28. Bornmann L, Schier H, Marx W, Daniel HD. Is interactive open access publishing able to identify high-impact submissions? A study on the predictive validity of Atmospheric Chemistry and Physics by using percentile rank classes. *J Assoc Inf Sci Technol* 2011; **62**: 61.
 29. Vanclay JK. Factors affecting citation rates in environmental science. *J Informetr* 2013; **7**: 265.
 30. Ketcham CM, Crawford JM. The impact of review articles. *Lab Invest* 2007; **87**: 1174.
 31. Mingers J, Leydesdorff L. A review of theory and practice in scientometrics. *Eur J Oper Res* 2015; **246**: 1.
 32. Tort AB, Targino ZH, Amaral OB. Rising publication delays inflate journal impact factors. *PLoS One* 2012; **7**: e53374.
 33. Martin BR. Editors' JIF-boosting stratagems – which are appropriate and which not? *Res Policy* 2016; **45**: 1.
 34. Aksnes D. A macro study of self-citation. *Scientometrics* 2003; **56**: 235.
 35. Van Raan AF. Sleeping beauties in science. *Scientometrics* 2004; **59**: 467.