Evidence-based recommendations for increasing the citation frequency of original articles

Short title: How to get your article cited

Nicole Heßler[a], Andreas Ziegler[b,c,d,*]

[a]Institut für Medizinische Biometrie und Statistik (IMBS), Universität zu Lübeck, Universitätsklinikum-Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

[b]Medizincampus Davos, Davos, Switzerland

[c]School of Mathematics, Statistics and Computer Science, University of KwaZulu Natal, Pietermaritzburg, South Africa

[d]Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[*]**Corresponding author**

E-mail: ziegler.lit@mailbox.org

ORCID:

Nicole Heßler: 0000-0003-0796-8172

Andreas Ziegler: 0000-0002-8386-5397

Abstract

Background: Publications and citations are important for career advancement of researchers. Our main aim was to derive recommendations that might increase the citation frequency of authors' work.

Methods: We examined title and article characteristics of original research articles published in the major medical journals BMJ, JAMA, Lancet, NEJM and PLOS Med (PLOS) between 2011-2020, using PubMed and Web of Science. To analyze citation frequencies, we estimated quasi Poisson regression models.

Results: The NEJM had by far the shortest titles (9.7±1.8 words). Titles in the other journals were at least 8 words longer on average. Randomized controlled trials (RCTs) were rarely identifiable by its title in the NEJM (5.3% by title, 63.3% by title plus abstract). BMJ, Lancet and PLOS articles had more frequently active verbs than JAMA and NEJM articles. The citation frequency was higher when articles were open access and when more authors and corporate authors were involved (all $p < 0.001$), and it was lower when a geography was mentioned ($p < 0.001$).

Conclusion: Titles differed substantially in their characteristics between major medical journals. The NEJM often chose titles for RCTs not following the CONSORT 2010 statement. Several modifiable title and article characteristics are associated with the citation frequency of articles, such as open access of an article. We recommend authors to choose the title carefully to obtain the maximum range for their work.

Keywords: Article title · Bibliometrics · Citation frequency · Impact · Research evaluation

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests: There are no patents, products in development or marketed products to declare. AZ is a licensed Tim Albert trainer and has held several courses in the past based on Albert's concept.

Availability of data and material: All relevant data are within the manuscript and its Supporting Information files.

Code availability (software application or custom code): All relevant code is within the Supporting Information files.

Authors' contributions: NH: Formal analysis, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing

AZ: Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing

Introduction

Scientists need publications and citations for their career advancement, but citations also affect the prestige of a scientific journal (Morgan, 1983). However, it is not only the scientific activity of a researcher that is evaluated by citation counts. Citations also express the popularity of the research topic (Bornmann and Daniel, 2008), which, in turn, may be directly related to the success of grant applications in the same research area by the scientist. There may also be a direct financial benefit for researchers because many research institutions distribute money according to citation records. Finally, the number of citations is correlated with the impact or influence of scientists and awards (Bornmann and Daniel, 2008). It is therefore of interest to identify factors that affect the citation frequency of an article. Approximately 30 of these factors have been identified so far, and they can be grouped into three categories (Ale Ebrahim et al., 2013, Bornmann and Daniel, 2008, Tahamtan et al., 2016). First, there are journal-related factors, such as the impact factor, the scope of the journal, the number of journals in the research area or its language. While authors cannot influence the journal-related factors, they may select the journal to which they submit their work. Interest in the journal, its wide distribution and its listing in specific databases, such as Medline or the Web of Science (Jeong and Huh, 2014) then contribute to the citation probability of an article. The second category is related to the author(s), such as the number of co-authors, international and national collaboration, sex, age, ethnicity and funding. It is, however, questionable whether author-related factors are modifiable for a specific article. For example, the number of citations increases with the number of co-authors (Bornmann and Daniel, 2008, Vieira and Gomes, 2010). This may just be an indicator for collaborative research, for which it has been argued that researchers who are open to collaboration produce superior outputs that result in a higher impact (Figg et al., 2006, Vieira and Gomes, 2010).

Article-specific factors form the final category. They include the document type, such as original article, review article or letter, novelty and interest of the study subject or the study design, such as animal study, case report or randomized controlled trial. Age of the article is an obvious factor because the number of citations increases over time. Even more, a frequently cited publication will be cited more frequently in the future, and there is a linear relationship between the expected number of future citations and the current number (Cano and Lind, 1991). Longer articles, such as review articles or tutorials, have more content than shorter articles, and the number of pages or words of an article therefore correlates with the citation frequency (Bornmann and Daniel, 2008, Hamrick et al., 2010). Similarly, articles with a large number of references are cited more frequently than works with fewer references (Bornmann and Daniel, 2008, Vieira and Gomes, 2010). However, some journals, such as the Lancet have upper limits in the number of references according to the publication type.

The article-related factors listed so far are not or hardly modifiable by authors, but authors can influence other characteristics of an article. Especially the title has been discussed as a modifiable component of interest (Fox and Burns, 2015). In fact, when scientists search databases for references, the title plays a major role. It provides keywords and index terms in the databases, and the occurrence of a database search term in an article's title has a strong impact on the ranking of an article in Google Scholar (Beel and Gipp, 2009). Most importantly, the title gives the reader a one-line summary. If the title matches the search topic, is somewhat catchy and listed among top hits of a search, it is likely that the researcher identifies the title as matching the search and may take a step towards a citation of the work by reading the abstract. If the abstract retains the attraction of the article, the researcher might wish to read the article. Simple access to the full text of the article, e.g., by a free of charge download option as pdf, might therefore be of great importance for an article to be cited. The title of an article thus is the modifiable main entrance gate to its possible citation. The association between the number of citations and title properties, such as the length of the title or the use of punctuation marks in

the title, have thus been studied. Previous analyses on these factors gave heterogeneous results. Some authors found more citations for articles with longer titles (Guo et al., 2018, Habibzadeh and Yadollahie, 2010). Others found exactly the opposite result (Guo et al., 2018). Finally, some reports are available where no association between title length and the number of citations was observed (Falahati Qadimi Fumani et al., 2015, Fox and Burns, 2015). Another title feature of interest is the geography, such as the country name or the city name in the title, and it is expected that such articles get fewer citations because of a more regional than general interest (Abramo et al., 2016). Researchers also investigated the use of subtitles (Fox and Burns, 2015) in the titles, among others. One such approach to investigate differences in citation frequencies of single or multipart titles is to see whether presence of a colon, which is by the most frequently used separator of parts in titles, leads to a larger number of citations.

Goodman (2000) studied the use of active verbs in titles. Specifically, he showed that the number of titles using an active verb increased between 1970 and 1997, and he predicted that 4.4% of all clinical papers in 2010 were going to have an active verb in the title. However, he did not analyze the effect of active verbs on the number citations. We expect that active titles are catchier and are therefore cited more frequently than other articles.

Differences in title length between journals were analyzed before. Kerans et al. (2016) showed that the New England Journal of Medicine (NEJM) had significantly fewer words than the BMJ (BMJ). Albert (2016) reported that BMJ titles were longest and had the largest proportion of titles with more than 20 words. The BMJ had fewer articles with less than 10 title words compared to the Lancet and the NEJM. The NEJM reduced the average number of title words from 11.6 in 2003 over 9.9 in 2005 to 8.6 in 2015 (Kerans et al., 2016). Since many original articles in the main medical journals are concerned with randomized controlled trials (RCTs) and, according to the CONSORT 2010 statement, a RCT should be identifiable as such from its title (Schulz et al., 2010), we expected a large proportion of articles consisting in at least two parts with a subtitle, where the colon was used as separator. We specifically expected a

difference in title length for articles reporting/not reporting on a RCT. This difference should be largest for the NEJM because we expected the NEJM to have the shortest titles among all considered journals. In consequence, adding a word, such as "random", "randomized" or even "randomized controlled trial" in the title should lead to the largest relative change in the number of words and characters.

The primary aim of our work was to examine title properties that may have an effect on the citation frequency of an original research article published in one of five major medical journals, which are BMJ, The Journal of the American Medical Association (JAMA), The Lancet (Lancet), NEJM and PLOS Medicine (PLOS). Furthermore, we investigated the effect of other article characteristics such as article length or number of references on the citation frequency. Specific hypotheses are formulated at the beginning of the next section.

Methods

Hypotheses for the length of the title

- H_1: Journals differ with respect to title length for both the number of title words (H_{1a}) and title characters (H_{1b}).
- H_2: The NEJM is expected to have shorter titles compared to the other four journals for both the number of title words (H_{2a}) and title characters (H_{2b}).
- H_3: The BMJ is expected to have longer titles than the Lancet and the NEJM for both the number of title words (H_{3a}) and title characters (H_{3b}).
- H_4: The number of title words remains stable for the NEJM over time.
- H_5: The title length varies between RCTs and other studies for both the number of title words (H_{5a}) and title characters (H_{5b}).

Hypotheses for the citation frequency

The following hypotheses are ordered by their degree of modifiability, and we hypothesize that articles are cited more often if

- the title is shorter (H_6) for both the number of title words (H_{6a}) and title characters (H_{6b}).
- the title has an active verb (H_7).
- titles are divided into multiple parts by the colon as separator (H_8).
- the geography is NOT mentioned in the title (H_{9a}), the city is NOT mentioned in the title (H_{9b}) or the country is NOT mentioned in the title (H_{9c}).
- the article is available free of charge (H_{10}).
- the number of authors is large (H_{11}).
- there are corporate authors (H_{12}).
- the number of pages is large (H_{13}).
- the number of references is large (H_{14}).
- the study is a RCT (H_{15}).

Additional aims

- Goodman predicted that 4.4% of all clinical papers in 2010 were going to have an active verb in the title. Our aim was to estimate the proportion of articles with an active verb for the period 2011 to 2020 (A_1).
- One additional aim is to estimate the number of title characters and title words (A_2).
- The proportion of free of charge articles increases over time for all journals, except for PLOS, which is a free access journal (A_3).

Search in Medline and Web of Science

We examined the titles of all original research articles published in the five major medical journals BMJ, JAMA, Lancet, NEJM and PLOS. Original research articles may be identified by using journal section headings, such as “papers” for the BMJ, “original contributions” for JAMA, “articles” for the Lancet and “original articles” for the NEJM. However, these articles are characterized by structured abstracts following the introduction, methods, results and discussion (IMRAD) structure (Sollaci and Pereira, 2004). We analyzed articles finally

published between 2011 and 2020. The restriction to the publication year 2011 was used to allow for proper comparisons between journals because PLOS was founded in 2004 and reshaped in 2009 (PLoS Medicine Editors, 2009).

For the citation analysis, we looked at citation frequencies for articles published until end of 2018. The search for the number of citations was done on Jan 01, 2021 in the Web of Science for retrieving the number of citations per year after publication. Details of the search are described in supplement 1, section 1. Publications had to be finally published not later than end of 2018 so that citations could be considered for a two-year period until the end of 2020. The article search in Medline was performed on Jan 01, 2021, and the detailed description of this search is provided in supplement 1, section 2.

Evaluation of Medline and Web of Science search

The following variables were extracted from the Medline search: PubMed identifier (PMID), journal name, article title, author names, publication year, citation, PubMed Central identifier (PMCID) and digital object identifier (DOI). For downstream analysis, we used the title and the author names—the latter to derive the number of authors.

From the Web of Science, we extracted journal name, article title, author names, publication date and year, PMID, DOI, abstract, the number of references and the number of pages. In addition, we investigated the open access type of a journal article. There are four main open access types: gold, bronze, green published and green accepted. Open access articles have the gold status if they are either listed in the Directory of Open Access Journals or fall under the Creative Commons license. Bronze open access articles are free-to-read or public access articles on a publisher's website. Green published articles are hosted in their final published version on an institutional or subject-based repository, such as PubMed Central. Finally, the green accepted status refers to articles which are hosted on a repository in their final accepted version, which may not have been copyedited or typeset by the publisher. Both PMID and DOI were used to merge articles identified in Medline and the Web of Science.

For the citations identified in the Web of Science we stored journal name, author names and DOI. Corporate authors, such as the "Australian New Zealand Intensive" or the "Irish Critical Care Trials Grp" were identified, and the variable was recoded to the presence or absence of corporate authors. In addition, we recorded the total number of citations, the average number of citations per year and the number of citations for each year, i.e., for 2011 until 2020.

We investigated the number of characters including blanks and the number of words per title, and we also divided the titles into less than 10 title words and at least 20 title words. Following Goodman (2000), we looked at active verbs and focused on the following 12 verbs: absolute verbs ("prevents," "abolishes," and "eliminates"), relative verbs ("prolongs," "reduces," "improves," "predicts," "lessens," and "weakens") and "nounal" verbs ("increases," "decreases," and "causes"). Therefore, we investigated the number of titles with one of the 12 verbs used in a declarational or phrasal way. Furthermore, we searched for the term "random*" in the title and/or abstract of an article to identify RCTs and other studies. The number of authors of an article was determined by the provided author names. The use of the colon as the separator was extracted from the title. For geographical information, we used the R package *maps* to identify city and/or country names in the title.

Statistics

For each year and each journal, means and standard deviations (SD) were calculated for continuous outcomes, and absolute and relative frequencies were reported for categorical variables. Analyses were done overall and, if appropriate, per year and per journal. The DerSimonian and Laird (2015; DSL) approach was used to perform random effect (RE) meta-analyses. The DSL approach allows for variability in the variables of interest between journals and over time. We calculated pooled RE estimates and standard errors. In particular, the logit transformation was used for the meta-analyses of binary outcomes for estimating the pooled proportions (Lipsey and Wilson, 2001), and standard errors were not back-transformed. The hypotheses about the title length were investigated by linear mixed models with journal as fixed

effect (FE) and year as RE, and, if appropriate with the binary variable RCT as FE. For the identification of homogenous subgroups in post hoc analysis, we applied Tukey's multiple comparison tests. Across hypotheses, we did not correct for multiple testing. The hypotheses about the number of total citations were investigated using quasi Poisson regression models, as recommended by Baggio et al. (2018). For sensitivity analysis, we employed quantile regression on the log number of total citations shifted by 1. Supplement 4 shows that the corresponding distributions were very symmetric. For all analyses, effect estimates and corresponding 95% confidence intervals (CIs) were estimated, and the BMJ was used as reference category. For the open access and corporate authors variables, analyses were performed with and without articles published in 2020 because the relevant information was completed in the Web of Science after our data export date.

Data and R code for all analyses are provided in Supplements 2 and 3, respectively.

Results

Descriptive statistics

The search for the five journals from 2011 including 2020 identified between 1313 articles (PLOS) and 2153 articles (NEJM). Table 1 provides an overview of the main characteristics for the articles from the five journals. Detailed results are provided in Supplement 4.

Substantial differences were observed for the proportion of articles that were classified as RCTs (Table 1). According to the CONSORT 2010 statement, RCTs should be identifiable in the title, and this is generally done with the word random* (Hopewell et al., 2020). While almost 2/3 (61.4%) of the articles published in the Lancet were classified as RCTs according to the title, only 5.3% of the articles published in the NEJM were categorized as RCTs by the databases. In JAMA, about every fourth article (42.2%) was a RCT, and it was about a fifth for BMJ and PLOS (21.0% and 18.8%, respectively). Frequencies changed substantially when the abstract

was used for identifying RCTs, and the biggest change was observed for the NEJM, where the frequency of RCTs increased to 63.3% (Table 1).

Another important difference was in the use of colons to separate the article title in two parts with a subtitle after the colon. Not a single colon was used among all NEJM titles (0 of 2153), while more than 95% of the articles published in the BMJ, the Lancet and PLOS had a two-part title with a colon. In 43.3% of the JAMA articles the title was divided in two parts with a colon.

Differences were also present in the mentioning of the geographical region, i.e., country or city, in the title. While the geography was part of the title in only 3% of the articles published in the US journals JAMA and NEJM, it was present in approximately 1/8 of the UK-based journals BMJ and Lancet (13.7% and 12.1%, respectively). The highest proportion of articles in which the geography was mentioned was PLOS with 31.9%, i.e., almost 1/3.

Few articles used active verbs in the title. The lowest percentages were observed for JAMA and the NEJM (1.4% and 2.0%, respectively), followed by the Lancet (3.9%) and PLOS (4.5%). The highest proportion of active verbs was noted for the BMJ (5.9%).

Some of the yearly results displayed in the supplement are worth mentioning. In our automated search we did not identify a single RCT in the NEJM in 2017 according to the title (Supplement 4, section 5.7.1). Furthermore, the proportion of RCTs varied substantially over the years for every journal (Supplement 4, section 5.7).

Hypotheses about title length

Descriptive statistics for title length of the journals are provided in Table 1. The number of characters of a title differed substantially between the journals. The Lancet had longest titles (22±6 words, 155±43 characters) and the NEJM by far shortest (10±2 words, 69±9 characters). BMJ and JAMA were closest in title length. The difference between these two journals was only 0.40 (95% CI: 0.07 – 0.74) words and 4.11 (95% CI: 1.78 – 6.44) characters, respectively, with JAMA having the slightly shorter titles. Because of the large sample size, all pairwise comparisons showed significant differences in title length for both the number of characters

and the number of words, except for the difference in the title words between BMJ and JAMA (adjusted p = 0.12).

The difference was also pronounced when the proportion of articles with < 10 title words and the proportion of articles with ≥ 20 title words were compared between the journals (Table 1; Supplement 4, sections 5.1.3 and 5.1.4). While almost a half (46%) of NEJM articles had < 10 title words, none of the NEJM articles had ≥ 20 title words. In contrast, more than half of the PLOS and the Lancet articles had ≥ 20 title words. Our first hypothesis (H_1) that journals differ with respect to title length was thus confirmed.

More specifically, we hypothesized that the NEJM had shortest titles (H_2), while the BMJ was expected to have longer titles than the Lancet and the NEJM (H_3). In fact, the NEJM had the shortest titles among all five journals. The average number of title words for the NEJM was below 10, while it was approximately 18 for BMJ and JAMA, respectively, and more than 20 for the Lancet and PLOS. The BMJ thus had longer titles than the NEJM ($p < 0.001$), but the Lancet had even longer titles than the BMJ ($p < 0.001$). Detailed results from regression models including 95% CI are given in Supplement 4, section 6.2.1. When we investigated the number of title words of the NEJM over time (H_4), we estimated an increase by 0.06 (95% CI: 0.03 – 0.08; $p < 0.001$) words per year (Table 2), which cumulated to 0.6 words over the 10-year observation period. In fact, the average number of words in NEJM titles was approximately 9.4 for the years 2011–2013, while it was approximately 9.9 for the years 2018–2020 (Supplement 4, section 5.1.2). In addition, there was an annual increase in the number of title words for all journals, which was significant for all journals ($p < 0.001$) but the BMJ ($p = 0.37$).

We expected differences in the title length for RCTs and other studies because randomized studies should be identifiable as such (H_5). Indeed, the title length was larger for RCTs by 2.21 words (95% CI: 2.00 – 2.42; $p < 0.001$) and by 19.44 characters (95% CI: 18.01 – 20.87; $p < 0.001$) compared to other original studies (Supplement 4, section 6.1.3.3 and section 6.2.3.3). The highest increase was observed for JAMA with about 6 (95% CI: 5.3 – 6.2) more

words in the title of a RCT, the lowest increase was for NEJM articles (mean difference 0.2 words, 95% CI: 0.0 – 0.3). Results for title characters are presented in Supplement 4 section 6.1.3.

Hypotheses about citation frequency

The number of citations was higher for longer titles (H_6). The risk ratio for the number of words was 1.02 (95% CI: 1.01 – 1.03; $p < 0.001$; Figure 1). Results from sensitivity analysis confirmed this finding ($p < 0.001$; Supplement 4, section 7.1.2.2). Results were similar but less pronounced when the title characters were investigated ($p = 0.08$ (Figure 1); sensitivity analysis $p = 0.02$ (Supplement 4, section 7.1.1.2)). The results were thus opposite to our expectation.

The proportion of articles with an active verb was reported in Table 1. Articles were more frequently cited when an active verb was used (risk ratio 1.23; 95% CI: 1.01 – 1.50; $p = 0.04$; Figure 1). However, the sensitivity analysis yielded a contradicting result for hypothesis H_7 (Supplement 4, section 7.2.2) because the association showed an effect in the opposite negative direction.

Hypothesis H_8 that articles are cited more frequently if the colon is used a separator of parts was not confirmed by our analyses. Both analyses yielded a non-significant difference (risk ratio 0.98; 95% CI; 0.85 – 1.13; $p = 0.78$; sensitivity analysis $p = 0.23$).

Mentioning the geography (H_9), i.e., city or country, in the title was associated with substantially lower citation frequencies (risk ratio 0.61; 95% CI: 0.51 – 0.73; $p < 0.001$; Figure 1; sensitivity analysis $p < 0.001$). Results were similar when either only mentioning of the city or of the country was considered (Supplement 4, sections 7.4.1 and 7.4.2).

A large difference in citation frequencies was observed between articles free of charge and other articles (H_{10}). Indeed, articles free of charge were cited substantially more frequently than articles that were not free of charge (risk ratio 1.43; 95% CI: 1.31 – 1.56; $p < 0.001$; Figure 1; sensitivity analysis $p = 0.02$; Supplement 4, section 7.5.2). Of note, the proportion of free access

articles increased over time. For years 2011 to 2019, the proportion increased by 1.4% per year (95% CI: 0.8 – 1.9; p = 0.002; Supplement 4, section 8.3).

Papers having more authors (H_{11}) were cited more often (risk ratio 1.02; 95% CI: 1.01 – 1.02; p < 0.001) as were articles having corporate authors (H_{12}; risk ratio 1.36; 95% CI: 1.27 – 1.46; p < 0.001). Sensitivity analyses confirmed these results (Supplement 4, sections 7.6.2 and 7.7.2).

A larger number of pages was associated with a higher number of citations (H_{13}), and the risk ratio was 1.04 (95% CI: 1.04 – 1.05; p < 0.001; Figure 1). Results were confirmed by the sensitivity analysis (p < 0.001; Supplement 4, section 7.8.2). Hypothesis H_{14} investigated whether a larger number of references was associated with a higher citation count. Both analyses clearly demonstrated the validity of this hypothesis (risk ratio 1.01; 95% CI; 1.01 – 1.01; p < 0.001; sensitivity analysis p < 0.001, Supplement 4 section 7.9).

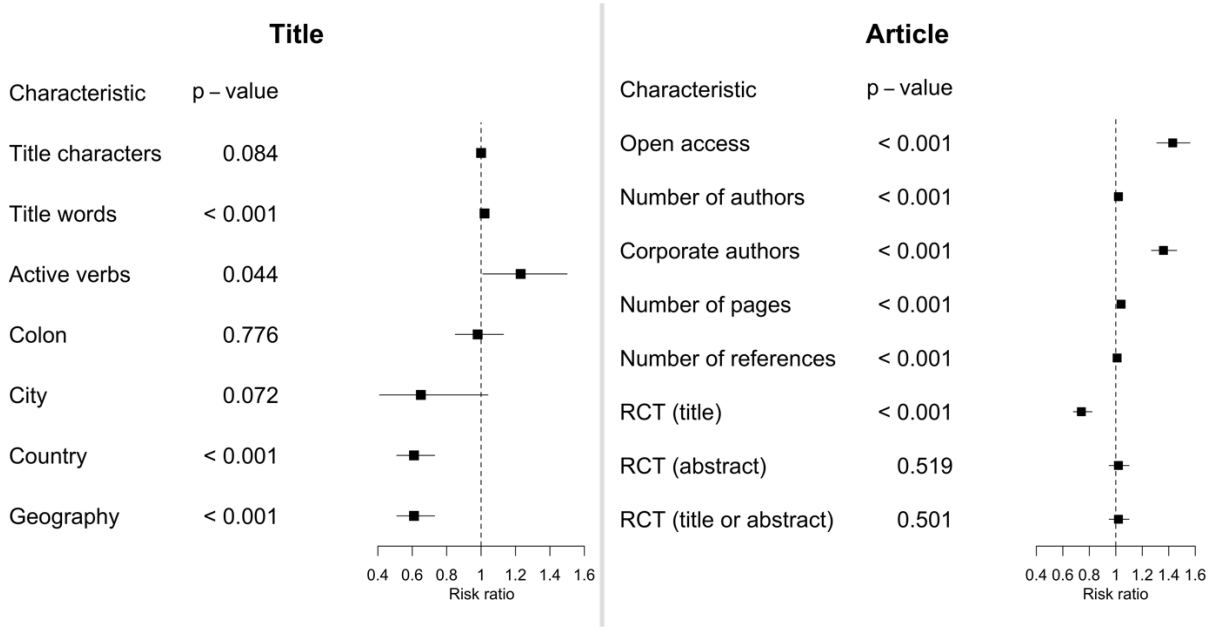

Fig. 1 Forest plots for variables possibly influencing citation frequency (left: title characteristics, right: article characteristics). Displayed are risk ratios and 95% confidence intervals. p-values are given as numbers. RCT = randomized controlled trial identified by term "random*" with * being the wildcard appeared in the title or in the title and/or abstract.

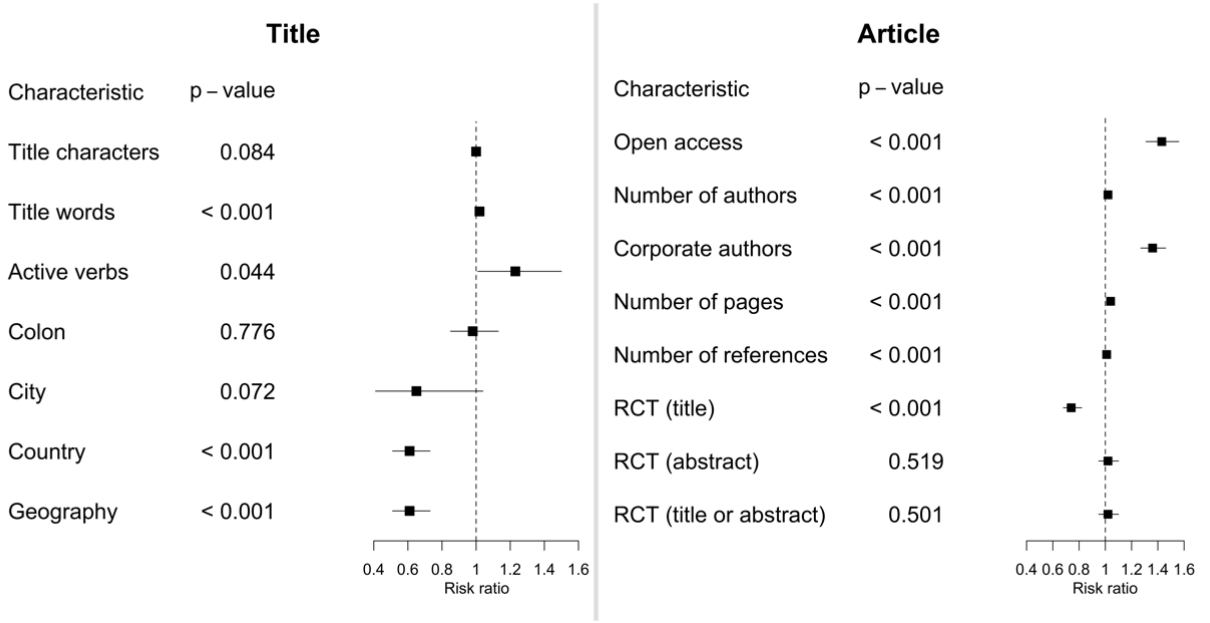

Fig.2 Forest plots for the title length for each journal (left: number of title characters, right: number of title words). Displayed are point estimates (β) and corresponding 95% confidence intervals (CI). Journal abbreviations are BMJ = The BMJ, JAMA = The Journal of the American Medical Association, Lancet = The Lancet, NEJM = The New England Journal of Medicine and PLOS = PLOS Medicine.

Finally, hypothesis H_{15} that RCTs are cited more frequently than other studies could not be shown by our primary analysis (risk ratio 1.02; 95% CI; 0.95 – 1.10; p = 0.50; but sensitivity analysis p = 0.03, Supplement 4 section 7.10.3.1 and section 7.10.3.2).

Additional analyses

Goodman (2000) predicted that 4.4% of all clinical papers in 2010 were going to have an active verb in the title. However, when averaged over the 10-year period from 2011 to 2020 and all journals, the proportion of articles with active verbs was 3.5% (95% CI: 3.0% – 4.2%, Supplement 4 section 8.1), thus lower than the percentage predicted by Goodman.

The second additional aim was the estimation of the number of title words and title characters. In addition to the descriptive statistics that were provided in Table 1, journal-specific estimates

from regression models plus 95% CIs are displayed in Figure 2, also see Supplement 4, sections 6.1.2.1 and 6.2.2.1.

The proportion of articles with open access increased over time for the Lancet from about 1/3 in 2011 and 2012 to more than 50% in last three years (Supplement 4, section 5.8). Results were similar for the NEJM, but here the proportion of open access articles increased from approximately 2/3 to more than 90% of the articles (Supplement 4, section 5.8). Overall, the proportion of open access articles increased over time from 86% in 2011 to 98% in 2019, pooled proportion of free of charge articles over all journals (except PLOS) and all years was 89 % (95% CI: 84.0% – 93%. Supplement 4 section 8.3).

Discussion

Title characteristics varied substantially between original articles published in top ranked medical journals from 2011 to 2020. Most striking were differences in title lengths. The NEJM had by far the shortest titles. Here, the average title length was below 10 words and no original NEJM article had a title with more than 20 words. Titles in the other four journals were at least 8 words longer on average than NEJM titles. Only few original articles had short titles, with the lowest percentage for PLOS, where only 2 out of 1313 (0.2%) original articles had short titles. Kerans et al. (2016) already reported that the NEJM had substantially shorter titles than the BMJ and two anesthesiology journals. However, the authors also reported a time trend towards shorter titles in the NEJM. Specifically, they reported that the NEJM reduced the average number of title words from 11.6 in 2003 over 9.9 in 2005 to 8.6 in 2015 (Kerans et al., 2016). In our trend analysis for original articles published in the years 2011 to 2020 we did not observe a decrease in title length in the NEJM. In contrast, the number of characters increased by 0.31 (95% CI: 0.18 – 0.44; $p < 0.001$) characters per year, corresponding to 0.06 (95% CI: 0.03 – 0.08; $p < 0.001$) words per year. Using data from 1995, Albert (2016) reported that BMJ titles were longer than Lancet and NEJM titles. Our analysis showed a different picture. Original

articles published in the BMJ from 2011 to 2020 were approximately 3.5 (95% CI: 3.2 – 3.9; $p < 0.001$) words shorter than articles published in the Lancet.

When authors aim to provide information about the study design in the title, this is generally done using a multi-part title, where parts are separated by a colon. Unexpectedly, not a single article published in the NEJM had a colon in the title. In contrast, almost all articles in the BMJ (99.0%), the Lancet (98.4%) and PLOS (97.2%) used colons in the title. Furthermore, we recognized that the NEJM did not fully comply with the CONSORT 2010 statement (Hopewell et al., 2020). Item 1a of the CONSORT 2010 checklist states that a randomized study should be identifiable by its title. In fact, a large proportion of RCTs published in the NEJM did not include any RCT identifier. Specifically, while only 5% of the articles were identified as RCTs in the NEJM by the title, approximately 60% of the articles were RCTs when the identification was extended to the abstract. This explains the unusual finding that not a single article in the NEJM was identified as RCT by its title in 2017 by the databases. However, when we searched the abstract for relevant keywords, the proportion of identified RCTs was very similar to values observed for years 2016 and 2018.

In agreement with Habibzadeh and Yadollahie (2010) and Guo et al. (2018) we observed higher citation frequency in original articles with longer titles. When we restricted the analysis to the subgroup of RCTs, a difference in the citation frequency of RCTs and other studies could not be detected (risk ratio: 0.98, 95% CI: 0.70 – 1.37, $p = 0.90$). This result may be surprising because RCT titles are on average 1.4 words longer than titles from other studies, however this can be explained by existing interactions with the factor journal.

The prediction of Goodman (2000) that 4.4% of all clinical papers in 2010 were going to have an active verb in the title was not correct. In fact, the proportion of articles with active verbs was only 3.5% (95% CI: 3.0% – 4.2%), when all journals and all years were considered, and it was just 2.6% (95% CI: 1.3% – 5.1%) for all journals in the year 2011. Interestingly, articles might be more frequently cited when an active verb is used in the title (risk ratio 1.23; 95% CI:

1.01 – 1.50; p = 0.04; Figure 1). However, the proportion of articles with active verbs was too low in this analysis for a definitive answer.

Although both the number of pages and the number of references were associated with higher citations (both $p < 0.001$), we are convinced that these two factors are hardly modifiable by the authors.

This is in contrast with articles being accessible free of charge and titles which mention the geography of the study. In line with our expectation, articles that can be accessed free of charge were cited more often than articles, which were not fully accessible to all potential readers (risk ratio 1.43; 95% CI: 1.31 – 1.56; $p < 0.001$). A title component associated with a lower number of citations was the mentioning of the geography, either in terms of the city or the country (risk ratio 0.61; 95% CI: 0.51 – 0.73; $p < 0.001$).

One limitation of our study is that we have investigated possible associations between citation frequencies and factors that are related to the study design or might be modifiable by the authors. However, only an experiment, which is impossible to conduct, could prove that modifications lead to higher citation frequencies. Another limitation of our study is that we relied on the quality of the data provided by two common databases, PubMed and Web of Science. In the data preparation phase, we identified several errors in these databases. For example, in one step we tried to identify original articles by structured abstracts with the PubMed search term "bmj" [Journal] AND "hasstructuredabstract" [All Fields] AND 2017 [dp]. While the number of articles with a structured abstract was between 150 and 160 for the years 2011 and 2020, it was just 29 for the year 2017, without a single article having a structured abstract between March and November. However, with the search term "bmj" [Journal] AND "hasabstract" [All Fields] AND 2017 [dp] we were able to identify articles with a structured abstract. When we looked at these articles in detail, we identified the abstract structure and noted that some formatting signs, such as end of line, feed forward seemed to be missing. In

our opinion, these relevant characters were not transmitted to the National Library of Medicine so that the abstracts were not categorized as hasstructuredabstract. We made the BMJ aware of this formatting issue (personal communication). However, we cannot exclude that we missed additional errors in the two databases. A third limitation of our work is that we were unable to study the effect of title content, such as the mentioning of Methods or Results, on the citation frequency (Kerans et al., 2020) as this requires the manual evaluation of titles. This is left for future research.

The results of this work are summarized in Box 1, which comprises a list of recommendations how the citation frequency of an original article might be increased.

Box 1. Recommendations how to increase the citation frequency of an original article; modified list by Norman (2012).

- Make your article open access. This simplifies potential readers to download and read on.
- Use an active verb in the title. This allows to transport the key message of the work.
- Make your article easy to find online. This recommendation has two components: first, choose the target journal for your article wisely as journals differ by the databases in which they are registered. Second, construct titles with keywords in mind. As pointed out by Norman (2012), search-unfriendly titles will not have much chance of getting seen.
- Avoid the use of multi-part titles. This will make the title catchier.
- Choose the length of the title according to the journal standard. This might increase the chance of acceptance by the editor (Albert, 2016).
- Avoid the use of geographic information.
- Disseminate your published article across your own networks, e.g., in presentations, and other media.

In conclusion, titles differ substantially in their characteristics between the five major medical journals BMJ, JAMA, Lancet, NEJM and PLOS. Unexpectedly, the NEJM often chose titles for RCTs that do not comply with the CONSORT 2010 statement. Several modifiable title and article characteristics are associated with the citation frequency of articles, such as free of charge access to the article. We recommend authors to choose the article title carefully to obtain the maximum range for their work.

References

Abramo, G., D'Angelo, C. A. & Di Costa, F. 2016. The effect of a country's name in the title of a publication on its visibility and citability. *Scientometrics,* 109**,** 1895-1909.

Albert, T. 2016. *Winning the Publications Game: The Smart Way to Write Your Paper and Get It Published,* Boca Raton, CRC Press.

Ale Ebrahim, N., Salehi, H., Amin Embi, M., Habibi Tanha, F., Gholizadeh, H., Motahar, S. M. & Ordi, A. 2013. Effective strategies for increasing citation frequency. *Int Educ Stud,* 6**,** 93-99.

Baggio, S., Iglesias, K. & Rousson, V. 2018. Modeling count data in the addiction field: Some simple recommendations. *Int J Methods Psychiatr Res,* 27**,** e1585.

Beel, J. & Gipp, B. 2009. Google Scholar's ranking algorithm: an introductory overview. *In:* Flory, A. & Collard, M. (eds.) *Proceedings of the IEEE International Conference on Research Challenges in Information Sciences, RCIS 2009.* Piscataway: IEEE.

Bornmann, L. & Daniel, H. D. 2008. What do citation counts measure? A review of studies on citing behavior. *J Doc,* 64**,** 45-80.

Cano, V. & Lind, N. C. 1991. Citation life cycles of ten citation classics. *Scientometrics,* 22**,** 297-312.

DerSimonian, R. & Laird, N. 2015. Meta-analysis in clinical trials revisited. *Contemp Clin Trials,* 45**,** 139-145.

Falahati Qadimi Fumani, M. R., Goltaji, M. & Parto, P. 2015. The impact of title length and punctuation marks on article citations. *Ann Libr Inf Stud,* 62**,** 126-132.

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C. & Birkinshaw, J. 2006. Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy,* 26**,** 759-767.

Fox, C. W. & Burns, C. S. 2015. The relationship between manuscript title structure and success: editorial decisions and citation performance for an ecological journal. *Ecol Evolut,* 5**,** 1970-1980.

Goodman, N. W. 2000. Survey of active verbs in the titles of clinical trial reports. *BMJ,* 320**,** 914-915.

Guo, F., Ma, C., Shi, Q. & Zong, Q. 2018. Succinct effect or informative effect: the relationship between title length and the number of citations. *Scientometrics,* 116**,** 1531-1539.

Habibzadeh, F. & Yadollahie, M. 2010. Are shorter article titles more attractive for citations? Cross-sectional study of 22 scientific journals. *Croat Med J,* 51**,** 165-170.

Hamrick, T. A., Fricker Jr., R. D. & Brown, G. G. 2010. Assessing what distinguishes highly cited from less-cited papers published in Interfaces. *Interfaces,* 40**,** 454-464.

Hopewell, S., Boutron, I. & Moher, D. 2020. CONSORT and Its Extensions for Reporting Clinical Trials. *In:* Piantadosi, S. & Meinert, C. L. (eds.) *Principles and Practice of Clinical Trials.* Cham: Springer International Publishing.

Jeong, G.-H. & Huh, S. 2014. Increase in frequency of citation by SCIE journals of non-Medline journals after listing in an open access full-text database. *Sci Ed,* 1**,** 24-26.

Kerans, M. E., Anne, M. & Sabaté, S. 2016. Content and phrasing in titles of original research and review articles in 2015: range of practice in four clinical journals. *Publications,* 4**,** 11.

Kerans, M. E., Marshall, J., Murray, A. & Sabaté, S. 2020. Research article title content and form in high-ranked international clinical medicine journals. *English Specif Purp,* 60**,** 127-139.

Lipsey, M. W. & Wilson, D. B. 2001. *Practical Meta-Analysis,* Thousand Oaks, Sage.

Morgan, P. P. 1983. The importance of being cited. *Can Med Assoc J,* 129**,** 9.

PLoS Medicine Editors 2009. A medical journal for the world's health priorities. *PLOS Med,* 6**,** e1000072.

Schulz, K. F., Altman, D. G., Moher, D. & for the CONSORT Group 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLOS Med,* 7**,** e1000251.

Sollaci, L. B. & Pereira, M. G. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc,* 92**,** 364-367.

Tahamtan, I., Safipour Afshar, A. & Ahamdzadeh, K. 2016. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics,* 107**,** 1195-1225.

Vieira, E. S. & Gomes, J. A. N. F. 2010. Citations to scientific articles: its distribution and dependence on the article features. *J Informetr,* 4**,** 1-13.

Tables

Table 1. Descriptive statistics for main characteristics of original articles for the period 2011 including 2020 (upper part) or 2019 (lower part).

Mean (standard deviation) are displayed for continuous variables, absolute and relative frequencies (parenthesis) for categorical variables. Journal abbreviations are BMJ = The BMJ, JAMA = The Journal of the American Medical Association, Lancet = The Lancet, NEJM = The New England Journal of Medicine and PLOS = PLOS Medicine. 1: Term "random*" with * being the wildcard appeared in the title or in the title and/or abstract. 2: proportion estimated after exclusion of unknowns. 3: Gold, bronze, green published or green accepted.

Characteristic	**BMJ (n = 1398)**	**JAMA (n = 1627)**	**Lancet (n = 1605)**	**NEJM (n = 2153)**	**PLOS (n = 1313)**
Title length					
Characters	130.3 (34.3)	126.6 (39.4)	154.6 (43.4)	69.4 (9.0)	140.2 (33.1)
Words	17.9 (4.8)	17.6 (5.6)	21.6 (6.2)	9.7 (1.8)	20.1 (4.9)
Less than 10 words	28 (2.0%)	88 (5.4%)	18 (1.1%)	988 (45.9%)	2 (0.2%)
At least 20 words	488 (34.9%)	569 (35.0%)	976 (60.8%)	0 (0.0%)	674 (51.3%)
Verbs					
Absolute	7 (0.5%)	0 (0.0%)	15 (0.9%)	28 (1.3%)	11 (0.8%)
Relative	36 (2.6%)	4 (0.2%)	23 (1.4%)	7 (0.3%)	23 (1.8%)
Nounal	40 (2.9%)	19 (1.2%)	26 (1.6%)	7 (0.3%)	26 (2.0%)
Active (sum)	83 (5.9%)	23 (1.4%)	63 (3.9%)	42 (2.0%)	59 (4.5%)
RCT[1]					
According to title	294 (21.0%)	688 (42.3%)	975 (60.7%)	115 (5.3%)	247 (18.8%)
According to abstract[2]	374 (30.6%)	815 (50.4%)	1039 (65.7%)	1336 (63.3%)	429 (33.1%)
According to title or abstract	397 (28.4%)	823 (50.6%)	1061 (66.1%)	1340 (62.2%)	434 (33.1%)
Geography	191 (13.7%)	64 (3.9%)	195 (12.1%)	66 (3.1%)	419 (31.9%)
Country	172 (12.3%)	56 (3.4%)	187 (11.7%)	60 (2.8%)	399 (30.4%)
City	29 (2.1%)	16 (1.0%)	23 (1.4%)	11 (0.5%)	55 (4.2%)

Colon	1384 (99.0%)	704 (43.3%)	1580 (98.4%)	0 (0.0%)	1276 (97.2%)
Analysis without year 2020	**(n = 1276)**	**(n = 1478)**	**(n = 1453)**	**(n = 1914)**	**(n = 1051)**
Corporate authors	135 (10.6%)	374 (25.3%)	570 (39.2%)	867 (45.3%)	121 (11.5%)
Open access[3]					
PubMed or Web of Science	1268 (99.4%)	1430 (96.8%)	642 (44.2%)	1537 (80.3%)	1051 (100.0%)
PubMed Central ID	1203 (94.3%)	881 (59.6%)	485 (33.4%)	772 (40.3%)	1051 (100.0%)

Table 2. Title length and development over time. Estimates and corresponding 95% confidence intervals (parenthesis) are displayed for each journal. Journal abbreviations are BMJ = The BMJ, JAMA = The Journal of the American Medical Association, Lancet = The Lancet, NEJM = The New England Journal of Medicine and PLOS = PLOS Medicine.

Characteristic	**BMJ**	**JAMA**	**Lancet**	**NEJM**	**PLOS**
Title length					
Characters	129.2 (126.2 – 132.3)	107.6 (104.3 – 110.9)	144.8 (140.8 – 148.8)	68.0 (67.3 – 68.7)	129.6 (125.7 – 133.6)
Words	17.8 (17.4 – 18.2)	14.8 (14.3 – 15.2)	20.4 (19.9 – 21.0)	9.4 (9.3 – 9.6)	18.5 (17.9 – 19.1)
Increase in length (per year)					
Characters	0.3 (-0.4 – 0.9)	4.4 (3.8 – 5.0)	2.1 (1.4 – 2.9)	0.3 (0.2 – 0.4)	1.9 (1.2 – 2.5)
Words	0.0 (-0.1 – 0.1)	0.7 (0.6 – 0.8)	0.3 (0.2 – 0.4)	0.1 (0.0 – 0.1)	0.3 (0.2 – 0.4)