Gender differences in citation sentiment: A case study in life sciences and biomedicine

Tahereh Dehdarirad (iD)
Department of Communication and Learning in Science, Chalmers University of Technology, Sweden

Maryam Yaghtin
Department of Scientometrics, Islamic World Science Citation Center (ISC), Iran

Abstract

In this study, we investigated whether female and male authors in the field of life sciences and biomedicine differed in their tendency for citation and citation sentiment. The data comprised two sets, cited set and citing set. Cited set comprised 17,237 articles whereas citing set comprised 115,935 articles. The cited set which is from the area Life Sciences & Biomedicine and published during 2012–2016 was retrieved from the Web of Science Medline. The citing set and its citation contexts were retrieved using the Colil database. The analysis was done using a combination of homophily analysis, regression analysis and a chi-square test. The covariates in the regression analyses were features related to authors, journal, institution, country and abstract readability. The homophily analysis showed a significant tendency for female (8%) and male (14%) authorship teams to cite papers by the same gender composition teams. In addition, the results of regression analysis (Model 1) and pairwise comparisons showed that male-authored papers received a significant higher positive sentiment compared with female-authored papers. The results of regression analysis (Model 2) showed a small significant positive association between gender similarity of cited and citing authorship teams and the sentiment score. However, further analysis using the chi-square test showed a significant lower tendency for women to use positive terms when citing the research findings of papers with the same gender composition. Men, in contrast, used significantly more positive terms when citing papers with the same gender composition. Finally, lay summary for a cited paper, country similarity and the venue of cited publication when it was a mega journal had a positive significant association with the sentiment score received.

1. Introduction

Women are under-represented in top-ranking positions in academia, such as professorships. This is true in the biomedical disciplines [1,2]. The way in which women present their research accomplishments, in comparison to men, has been regarded as one of the factors that causes gender gaps in academic life sciences according to previous research [3]. A group of studies on gender and self-promotion have investigated the association between gender gap/implicit bias and language patterns. Lerchenmueller et al. [3] analysed the language used in more than 6 million papers in peer-reviewed articles in life sciences. Their findings showed that papers with a male first author or a male last author were up to 21% more probably to use positive framing than those with female lead authors. In mathematics, it has been shown that women are less probably to boast about their intellectual performance [4]. In a study on gender gaps in grant funding at Canadian Institutes of Health Research, Witteman et al. [5] associated the gender gap in grant funding to the unconscious use of modest, less compelling language by female applicants, when describing their accomplishments. However, in a more recent study by Franco et al. [6], no significant gender differences were found in terms of language patterns

Corresponding author:
Tahereh Dehdarirad, Department of Communication and Learning in Science, Chalmers University of Technology, Hörsalsvägen 2, 41296 Gothenburg, Sweden.
Email: Tahereh.Dehdarirad@chalmers.se

used in biomedical research grant proposal summaries in Brazil. Lewis and Lupyan [7] showed that language statistics can predict people's implicit biases. In other words, linguistic biases may be causally related to biases in people's implicit judgements of what can be accomplished by women.

As the notion of 'excellence' within audit cultures in academia implicitly reflects images of masculinity [8], the promotion of women's research findings could be considered counter stereotypical behaviour, which may in turn cause a backlash against women [9–11]. Men's research can be viewed as more central or important in a field (i.e. the 'Matthew' effect), whereas women's work can be ignored, or worse, attributed to men in a field (i.e. the 'Matilda' effect) [12,13]. This can, therefore, result in implicit biases in term of authors' tendency regarding citations [14] and citation sentiments. In a large-scale study of gender citation homophily, Ghiasi et al. [15] suggested under-recognition of women's contribution to science, as a reason for men's preference for citing men.

In this article, we investigated when citing a paper, whether female and male authors differed in terms of their citation tendency and framing the findings of papers that they cited, using inbreeding homophily and sentiment analyses. To the best of our knowledge, no study so far has investigated this in the field of life science and biomedicine. Inbreeding homophily refers to the ties between similar others that form over and beyond what would be expected by chance given the opportunity pool [16]. Citation sentiment analysis is a natural language processing (NLP) technique that has been used in previous research to study the sentiment of an author towards a paper that they have cited [17–20]. Citation content analysis research focuses on describing and classifying semantic relationships between the citing and cited works [21], providing a more nuanced, qualitative view of how an article is cited [22]. These classifications are based on the polarities and strengths of the opinions extracted from citances[1] [23]. In the literature, two different methods have been used for classification of citation sentiments: unsupervised (Lexicon-based) approaches and supervised (machine learning framework) approaches [17]. Most unsupervised methods rely on a sentiment lexicon to assign a sentiment score to the text. This unsupervised approach is more understandable and can easily be implemented in contrast to machine learning-based algorithms [24]. However, sentiment lexicons are largely dependent on the genre of the text being classified [17]. Studies [25-27] are example research works in life sciences and biomedince that are conducted using Lexicons. This method has also been applied in citation sentiment analysis studies in the context of co-citation analysis [28], to rank citations [29] or to measure citation quality [30]. The supervised methods represent the labelled data in the form of a set of features. The features are then used to learn a function for classification of unseen data [17]. While one advantage of learning-based methods is their ability to adapt and create trained models for specific purposes and contexts, their drawback is the availability of labelled data and hence the low applicability of the method on new data. This is because labelling data might be costly [31]. Using supervised machine learning techniques, some studies in biomedicine have been conducted on small, annotated citation corpuses from biomedical articles [32,33] or clinical trials [18,22].

This study has two objectives: (1) in general, to study whether there is a tendency towards authorship teams based on their gender, in terms of sentiment, when citing or being cited and (2) more specifically, to investigate whether there is a tendency for authorship teams, in terms of citations, sentiment score or use of positive/negative terms, when citing papers of the same gender composition teams.

To achieve these objectives, the study addressed the four below questions:

1. Is there any difference in terms of sentiment score among authorship teams of cited and citing papers, based on their gender?
2. Is there any citation homophily based on gender between citing and cited papers?
3. Is there any association between gender similarity of cited and citing authorship teams and the sentiment score received?
4. Is there any difference between authorship teams in terms of the proportion of positive/negative terms used, when citing papers of the same gender composition teams?

2. Methodology

2.1. Data sets

The data for this study comprised two sets, the cited set and the citing set. The cited set comprised 17,237 articles, whereas the citing set comprised 115,935 articles. By the citing set, we mean all of the articles that cited the 17,237 articles in the cited set.

To retrieve these data sets, several steps were taken. First, all articles in the research area of Life Sciences & Biomedicine from 2012 to 2016 were retrieved from the Web of Science Medline in June 2020. This resulted in 83,444 papers. In the next step, using the PubMed identifier (PMID) of these papers, a search was conducted in the PubMed

Central Open Access Subset (PMC-OAS). As a result of this search, we identified 17,237 papers in PMC-OAS. The PMC-OAS is a part of the total collection of articles in PMC, which are made available under a Creative Commons or similar licence that generally provides users with the rights to reuse and redistribute content [34]. In the next step, citations and citation contexts for each of these 17,237 cited papers were extracted from the Colil (Comments on Literature in Literature) database. This resulted in 115,935 citing articles, which provided 210,181 citation contexts. These two data sets, citing and cited, were used for further analysis in this article.

Colil is a database and search service for citation contexts in the life sciences. Using the web-based search service, it is possible to search for a cited paper in the Colil database and retrieve a list of its citation contexts. The citation contexts in the Colil database are extracted from full-text papers of the PMC-OAS [35].

2.2. Data analysis and procedures

2.2.1. Sentiment analysis. KNIME (the Konstanz Information Miner) was used to pre-process the citation contexts obtained from the data collection process. The pre-processing step included removing stop words, special characters and numbers. It also entailed tokenisation, part of speech (PoS) tagging and lemmatisation. In this article, sentiment analysis was conducted on the four parts of speech, namely, nouns, adjectives, adverbs and verbs. The reason for this was that these parts of speech are believed to carry opinion loads in a sentence [36,37]. We also accounted for three valence shifters: amplifiers (absolutely, certainly), de-amplifiers (barely) and adversative conjunctions (although). In the case of negators (not, can't, etc.), we applied a list of negation terms[2] that was used in Athar's [17] study. In addition, we also added other negators such as aren't, didn't and has not to the list. The number of negators was very low (~ 0.5%) in our sample.

The sentiment scores were assigned to words using SentiWordNet. SentiWordNet is one of the most widely used sentiment lexicons [38] which has a wide coverage of words [39]. Furthermore, it has been previously used in the domains of medicine and biomedicine [25,26] and in citation sentiment analysis studies [29,30,40].

SentiWordNet assigns three sentiment scores to each synset of WordNet: positivity, negativity and objectivity, ranging from 0 to 1 [38]. Since every word may have different meanings, SentiWordNet has a collection of different synsets.[3] Every synset may have positive, negative and neutral scores for a given word, which sums up to one. Given the existence of several synsets for a given word, it was necessary to assign a single score to each word. To do so, we used a methodology applied by Cavalcanti et al. [29]. For a given word, within a specific PoS, all scores for its synsets were averaged, after subtracting the negative scores from the positive ones. Finally, the scores of each citation context were aggregated. In some cases, a citing article might have referenced a cited article in several places: thus, creating more than one citation context. In these cases, the sentiment score of these citation contexts was averaged.

Finally, we randomly selected 5% (~850) of the cited PMIDs. For each of these PMIDs, we then randomly selected one citing PMID and one of its corresponding citation contexts. The minimum required sample size for inter-rater agreement in this study was calculated using the methodology provided in Rotondi and Donner [41]. These contexts were then read by two experts in the field of life sciences and biomedicine, and they assigned sentiment polarity to the sample citation contexts. To measure the inter-rater agreement between each person's polarity and the automated polarity, we used both percent agreement and Cohen's kappa test using the *irr* [42] package. The latter is used to measure inter-rater reliability for qualitative (categorical) items [43]. Values range between -1 and 1. Values lower than 0 are indicative of no reliability, values between 0 and 0.4 of slight to fair, values between 0.41 and 0.6 of moderate, values between 0.61 and 0.8 of substantial and values between 0.81 and 1 of almost perfect reliability [44]. In our case, the results showed a kappa agreement of 0.68 and 0.70 (~70%), which indicates a good agreement. In addition, the percent agreement showed values of 79% and 82%. As the experts were from the field of life sciences and biomedicine, they would be less probably to guess when assigning sentiments. Thus, we believe these percent values are also reliable [43].

2.2.2. Regression and homophily analysis. Regarding questions one and three, the linear mixed model (LMM) was used to: (1) to compare the mean of sentiment scores by gender among authorship teams associated with cited and citing articles; (F (both first and last female), M (both first and last male), M-F (first male-last female), F-M (first female-last male)) (Model 1) and (2) to study the association between sentiment score and gender similarity (measured as cosine similarity between cited and citing sets based on gender), while controlling for some covariates (Model 2) (see Table 1). Using Model 2, we investigated whether there was a tendency for authors (not) to be more positive towards the authors of similar gender composition in terms of terms of sentiment score.

LMM is a generalisation of linear regression, which allows to model correlated errors. These models are called mixed effects since they combine both fixed effects and random effects. While using fixed effects, we can get an estimate of their effects on the response variable (in our case sentiment score), using random effects, we can quantify the variability

Table 1. Dependant, independent variables and covariates for regression Models 1 and 2.

Effect type	Variable type	Name	Measure
Predictors (fixed effect)	Dependent	Sentiment score	
	Independent	Citing authorship teams by gender[a]	Gender combination of first and last author on a citing article: M (both authors male), F (both authors female), F-M (first author female, last author male), M-F (first author male, last author female).
		Cited authorship teams by gender[a]	Gender combination of first and last author on a cited article: M (both authors male), F (both authors female), F-M (first author female, last author male), M-F (first author male, last author female).
		Gender cosine similarity[b]	Cosine similarity between cited and citing sets based on gender of first and last authors.
	Covariate	SNIP-cited set	Source (journal) Normalised Impact per Paper for the cited articles which was downloaded from Scopus.
		Subject cosine similarity	Cosine similarity between cited and citing sets based on MeSH.
		Lay summary-cited set	Cited articles with lay summary (1); without lay summary (0).
		Abstract readability-cited set	Flesch's readability score of the abstracts of cited articles.
		Country cosine similarity	Cosine similarity between cited and citing sets based on affiliation country of all authors.
		Affiliation country of cited author; affiliation country of citing author	For each of cited and citing sets, a variable was created based on the affiliation country of all authors: cited English-speaking country (yes); non-English-speaking country (no), citing English-speaking country (yes); non-English-speaking country (no).
		Mega journal-cited set	If journals associated with cited set were: mega journal (1) or non-mega journal (0).
		OA-cited set	If cited articles were OA (1) or non-OA articles (0).
		Scientific age of citing authors	The total number of citations for all authors associated with each article in the citing set were obtained from SciVal author look up API (https://dev.elsevier.com/documentation/SciValAuthorAPI.wadl) using authors' IDs. Then, geo mean of citations was calculated for each citing article.
		Top institute	A variable which showed whether a top rank university existed in the affiliation list of cited and citing articles: if existed on none (0), if existed on both (1), if existed on cited but not citing (2), if existed on citing but not cited (3).
Predictor (random effect)		Cited PMID	It was entered in both regression models (Model 1 and Model 2) as random = ~1|Cited PMID

SNIP: Source Normalised Impact per Paper; OA: open access; API: application programming interface; PMID: PubMed identifier.
[a]These were the independent variables in Model 1.
[b]This was the independent variable in Model 2.

in the baseline values of the response variable [45,46]. Mixed-effect models are the preferred approach for the analysis of longitudinal or clustered data [47,48]. The reason for this is flexibility of these models in modelling the correlation structure of the data and ability to account for both balanced (i.e. with the same set of observations points for each subject) and unbalanced (i.e. with different observations points for different subjects) designs. Another major advantage of LMM is how they handle missing data. Specifically, the mixed-effects approach does not require a complete set of observations on each subject, rather, it uses all available data on each subject. If a subject has intermittent missing data, or subjects/clusters that have different numbers of (repeated) observations, all available data are used [48]. The case in this study is clustered and unbalanced data design. The grouping variable is the cited articles' PMIDs. All citing articles (citing PMIDs) are matched with their cited PMIDs. As the clusters (cited PMIDs) can be considered to have been sampled from a larger population of clusters, their effects can be modelled as random effects in an LMM [47]. Furthermore, the sentiment scores within cited PMIDs might be correlated. The reason for this is that we expect the citing PMIDs to be more homogeneous within a cited PMID, than between cited PMIDs.

In both regression models, the outcome y (sentiment scores) is a continuous variable. Furthermore, we have 13 fixed effects predictors with fixed slopes for Model 1: gender citing, gender cited, top institute, subject similarity, open access (OA) status, abstract readability, SNIP (Source Normalised Impact per Paper), lay summary, scientific age of citing authors, country similarity, affiliation country of cited author (English-speaking country or not), affiliation country of citing author (English-speaking country or not) and mega journal (see Table 1). In Model 2, we have the same fixed

effect predictors except for gender-citing and gender-cited variables instead of which we used gender cosine similarity. In both models, we have a random effect associated with the intercepts for each cited PMID. As mentioned above, random effect is a way to correct statistical tests when some observations are correlated [37]. We fitted the LMM using lme () function from the R package *nlme* [49]. R package *Emmeans* [50] was also used to perform pairwise comparisons among authorship teams of different gender (F, M, M-F, F-M) associated with cited and citing articles.

When fitting the regression models: (1) we tested whether the random effects associated with the intercept for each cited PMID can be omitted from the regression Models (1, 2) and (2) we also checked to decide which covariates to include in the model. To do so, for both cases, we conducted a likelihood ratio test using the anova () function. In the first case, the result of the test ($p < 0.001$) for both Models (1, 2) suggested that the random-cited PMID effects should be retained in the regression Models (1, 2). Regarding covariates, after running the anova () function on the random effect models for both Models 1 and 2, we decided to remove the scientific age of cited authors from both models as it did not improve the models. The *sjPlot* R package [51] was used for diagnostic of both models in terms of multicollinearity, normality of residuals, normality of random effect and homoscedasticity (see Supplementary Material).

To answer the second question, we investigated whether there was an inbreeding homophily in citation tendency by gender composition of authorship teams (cited-citing). To do this, we used inbreeding homophily analysis using R.[4]

Finally, to answer question 4, a chi-squared test was used to compare the proportion of positive and negative terms between the same gender authorship teams associated with both cited and citing sets (both female teams, both male teams, both F-M teams, both M-F teams).

2.2.3. Data collection and processing of the variables. In this section, we provide details regarding data collection and processing of some of the variables used in the regression analysis. We also provide reasons why we have included these variables as covariates in the regression models.

Several of the covariates used in the regression Models (1, 2) capture different aspects that might have an impact on the sentiments towards a cited work based on the previous literature. Journal aspects (SNIP and mega journal): as higher impact journals might publish more papers of higher importance, they tend to be cited with more positive sentiment [20]. Readability aspects (lay summary and abstract readability): previous research has found that easy to read posts are more liked, commented on and shared on social media [52]. Furthermore, the effect of language sentiment is more probably when readability is low than when it is high [53]. So, we expect text readability to have a positive influence on the sentiments and attitude towards cited works. Institute's and author's reputation (for which the proxies were top institute and scientific age): as reputation can be considered as important signal of trustworthiness and quality [54], it is expected that papers with authors and institutes with higher reputations to receive more positive sentiments. Furthermore, previous research has shown a relationship between author's and institute's prestige and the number of citations received [54]. Country-related aspects (measured as country similarity and author's affiliation with a country where English is the official majority language (Australia, New Zealand, United Kingdom, Canada and the United States) and authors with affiliations outside these countries). Previous research has shown differences regarding the use of positive words between authors affiliated with these two groups of authors [55].

To detect the gender of first and last authors in both cited and citing sets, we used a combination of gender application programming interface (API) (https://gender-api.com/) and manual checking. First, gender API was used to conduct a search using the first name of authors. For each name, the API returns the gender (male, female or unknown), the number of names used to determine the gender and accuracy [56]. Then, in cases of gender-neutral, unknown, initials or where the accuracy was lower than 80%, the names were checked manually using Internet searches. The reason for choosing these two authorship positions was that in the field of biomedicine, the last position in the authors list is reserved for senior authors, whereas the first author position is for the person who fulfils the International Committee of Medical Journal Editors (ICMJE) authorship criteria to the highest level and performs the majority of the experimental and clinical work [57,58]. We calculated cosine gender similarity and cosine subject similarity between citing and cited sets using Python. The subject similarity was calculated based on MeSH. For each article in both cited and citing sets, MeSH was downloaded from Medline using *Bio.Medline* Python package [59]. MeSH pairs of citing/cited articles have been used in the previous literature for identifying shared medical topics between cited and citing articles and clustering of topically related articles in biomedical sciences [60,61].

The scientific age of authors on citing/cited papers was calculated using the geometric mean of citations[5] for first and last authors. The number or log-transformed number of publications and citations of an author has been previously defined as professional or scientific age of an author [62–64]. While we used the Web of Science Medline to determine the OA status of the cited articles, we used the list provided by Spezi et al. [65] to determine whether a journal was a mega journal or not.
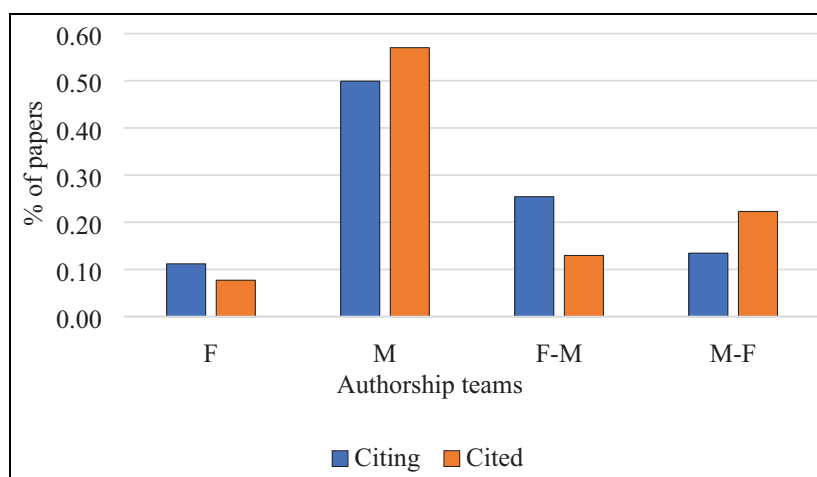
Figure 1. The percentage of papers for authorship teams of different gender in cited and citing sets.

Table 2. The result of inbreeding homophily analysis for citation tendency by gender composition of authorship teams.

Authorship team	Estimates	Significance
F	0.083	$< \mathbf{0.0001}$***
F-M	0.00001	0.5
M	0.142	$< \mathbf{0.0001}$***
M-F	0.00001	0.5

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Regarding readability we calculated 'abstract readability' and 'lay summary[6]' for each cited article. Using a journal list provided by Shailes [66], the cited articles were divided in two groups – those with and those without a lay summary. Regarding abstract readability, the Flesch Reading Ease Score was used, as it is the most used measure of text readability, and it has been used in other bibliometric studies [67]. The R *quanteda* package [68] was used to calculate this score for each abstract. The highest possible score is 121.22 and there is no lower limit. Very complicated sentences can have negative scores. The higher the score, the easier the text is to understand.

3. Results

Figure 1 shows the percentage of papers for citing and cited authorships teams by gender. As can be seen from the figure, in both sets, male authorship team was the group with the highest number of papers in both cited and citing sets, whereas female authorship team was the group with the lowest number of papers in both sets.

The results of homophily analysis showed that there was a significant inbreeding for female and male authorship teams. This means that there is a significant tendency for female and male authorship teams to cite the papers by the same gender composition teams. As can be seen from Table 2, 14% of the times an authorship team composed entirely of male authors will cite another authorship team composed of only male authors. For female authorship teams, this means that 8% of the times, a female authorship team will cite another female authorship team.

3.1. The results of linear mixed-effect regression analysis for Models 1 and 2

In this section, we have reported the results of regression analysis regarding Model 1 and Model 2, while controlling for several covariates. In the tables for both models, ICC is the intraclass correlation coefficient, which is calculated by dividing between-variance and the sum of between- and within-variance. In other words, ICC = $\tau_{00}/(\sigma^2 + \tau_{00})$, where σ^2 is the between-group variance and τ_{00} is equal to within-group variance [69]. By looking at the ICC values for both

Table 3. The results of linear mixed-effect analysis for Model 1.

Predictors	Sentiment score			
	Estimates	Exp.	CI	p
(Intercept)	1.88	6.55	1.66 to 2.10	< **0.001**
Gender-citing (F-M)[a]	0.02	1.02	− 0.00 to 0.05	0.088
Gender-citing (M)	0.04	1.04	0.01 to 0.07	**0.002**
Gender-citing (M-F)	0.04	1.04	0.01 to 0.07	**0.007**
Gender-cited (F-M)[b]	0.03	1.03	− 0.03 to 0.09	0.355
Gender-cited (M)	0.04	1.04	− 0.01 to 0.09	0.145
Gender-cited (M-F)	− 0.04	0.96	− 0.10 to 0.01	0.132
SNIP-cited	0.00	1.00	− 0.01 to 0.02	0.726
Lay summary-cited (Yes)	0.24	1.27	0.19 to 0.28	< **0.001**
OA-cited (Yes)	0.21	1.23	0.00 to 0.42	0.050
Citing authors-scientific impact	− 0.01	0.99	− 0.01 to − 0.00	**0.001**
Top institute (1)	− 0.03	0.97	− 0.06 to 0.00	0.080
Top institute (2)	− 0.02	0.98	− 0.04 to 0.01	0.137
Top institute (3)	− 0.00	1.00	− 0.04 to 0.03	0.768
Subject similarity	0.07	1.07	− 0.03 to 0.16	0.174
Cited-mega journal (yes)	0.03	1.03	0.00 to 0.06	**0.035**
Country similarity	0.03	1.03	0.01 to 0.05	**0.010**
Cited_English-speaking country (no)	− 0.06	0.94	− 0.09 to − 0.02	< **0.001**
Citing_English-speaking country (no)	− 0.03	0.97	− 0.04 to − 0.01	**0.004**
Cited-abstract readability	0.00	1	− 0.00 to 0.00	0.227
Random effects				
σ^2	1.58			
τ_{00} cited PMID	0.40			
ICC	0.20			
Number of cited PMIDs	17,237			
Observations	115,935			
R^2	0.58			

CI: confidence interval; SNIP: Source Normalised Impact per Paper; OA: open access; PMID: PubMed identifier; ICC: intraclass correlation coefficient.

[a]Gender citing (F) as the reference group.

[b]Gender cited (F) as the reference group.

Table 4. Results of pairwise comparison based on gender among cited authorship teams (gender cited) and among citing authorship teams (gender citing).

Contrasts: authorship teams by gender	Cited set			Citing set		
	Estimates	SE	Significance	Estimates	SE	Significance
F − (F-M)	− 0.030	0.032	0.791	− 0.024	0.014	0.321
F − M	− 0.039	0.027	0.464	− 0.040	0.013	**0.010**
F − (M-F)	0.044	0.029	0.434	− 0.042	0.016	**0.036**
(F-M) − M	− 0.010	0.022	0.970	− 0.017	0.009	0.291
(F-M) − (M-F)	0.074	0.025	**0.015**	− 0.018	0.013	0.499
M − (M-F)	0.083	0.018	< **0.0001**	− 0.001	0.012	0.999

SE: standard error.

Models 1 and 2, it can be said that 20% of the variance in sentiment scores is due to variation between cited articles, whereas 80% of the variance is due to variation within cited articles (among citing articles connected to a cited article). Regarding gender, the result of regression analysis (Model 1) (Table 3) and pairwise comparisons (Table 4), showed that (1) in citing teams, female only authorship (F) teams had a significant lower positive sentiment towards the cited papers in comparison to male (M) and mixed (M-F) authorship teams. This means that female authorship teams were slightly

Table 5. The results of linear mixed-effect analysis for Model 2.

Predictors	Sentiment score			
	Estimates	Exp.	CI	p
(Intercept)	1.89	6.62	1.68 to 2.10	< **0.001**
Gender similarity	0.04	1.04	0.01 to 0.07	**0.004**
SNIP-cited	0.00	1.00	− 0.01 to 0.02	0.631
Lay summary-cited (yes)	0.24	1.27	0.20 to 0.29	< **0.001**
OA-cited (yes)	0.21	1.23	0.00 to 0.42	0.050
Citing authors-scientific impact	− 0.00	1.00	− 0.01 to − 0.00	**0.001**
Top institute (1)	− 0.03	0.97	− 0.06 to 0.00	0.058
Top institute (2)	− 0.02	0.98	− 0.04 to 0.01	0.139
Top institute (3)	− 0.01	0.99	− 0.04 to 0.03	0.685
Subject similarity	0.07	1.07	− 0.03 to 0.16	0.187
Cited-mega journal (yes)	0.03	1.03	0.00 to 0.06	**0.036**
Country similarity	0.03	1.03	0.01 to 0.05	**0.013**
Cited_English speaking country (no)	− 0.06	0.94	− 0.09 to − 0.03	< **0.001**
Citing_English speaking country (no)	− 0.03	0.97	− 0.04 to − 0.01	**0.004**
Cited-abstract readability	0.00	1	− 0.00 to 0.00	0.232
Random effects				
σ^2	1.58			
$\tau 00$ cited PMID	0.40			
ICC	0.20			
Number of cited PMIDs	17,237			
Observations	115,935			
R^2	0.52			

CI: confidence interval; SNIP: Source Normalised Impact per Paper; OA: open access; PMID: PubMed identifier; ICC: intraclass correlation coefficient.

less positive towards cited papers in comparison to male (M) and mixed authorship teams (M-F) and (2) in cited teams, male teams (M) received a significant higher positive sentiment in comparison to mixed teams (M-F). In addition, mixed authorship teams with a female last author (M-F) received a lower positive sentiment in comparison to mixed authorship teams with a male last author (F-M).

The result of regression analysis regarding gender for Model 2 (Table 5) showed a small significant positive association between gender similarity of cited and citing authorship teams and sentiment score received per cited papers. This means that as the gender similarity between cited and citing sets increased, the sentiment score increased.

As for the rest of covariates in both models, the results showed that the existence of lay summary for a cited paper had a small positive significant association with the sentiment score received. In addition, the results showed a positive association between the publication venue of a cited paper when it was a mega journal and the sentiment score revived. This means that if the publication venue of a cited paper was a mega journal, the sentiment score received was positive.

Regarding the country covariates, as can be seen from Table 5, as country similarity increased, the sentiment score increased. In addition, the results showed when authors on citing or cited papers were affiliated with non-English native speaking countries, the sentiment score decreased.

3.2. The results of chi-square test to compare the proportion of positive and negative terms

A chi-square test was used to study whether the proportion of counts of positive and negative terms between the same gender composition teams was significantly different or not. The results showed that relative proportion of positive and negative terms was significantly different in the groups ($p < 0.0001$). In addition, the result of post hoc test (using Bonferroni's correction) showed that the proportion of positive terms was significantly higher than expected for male authorship teams when citing a male authorship team. Interestingly, the proportion of negative terms was significantly higher than expected for female and mixed gender (F-M and M-F) teams, when citing teams composed of the same gender.

4. Conclusion and discussion

In this study, we investigated whether female and male authors in the field of life sciences and biomedicine differed in their citation homophily and framing the findings of papers that they cited. To do this, we addressed two research objectives and four questions. The findings in relation to these are discussed below.

Regarding the first objective, whether there was a tendency towards authorship teams by gender in terms of sentiment when citing or being cited, we investigated this via question one. The results of regression analysis and pairwise comparisons for question 1 showed that when citing a paper, female only authorship (F) teams had a significant lower positive sentiment towards the cited papers in comparison to male (M) and mixed (M-F) authorship teams. In other words, female authorship teams were slightly less positive towards cited papers in comparison to male (M) and mixed authorship teams (M-F). In addition, when being cited, male teams (M) received a significant higher positive sentiment in comparison to mixed (M-F).

Regarding the second objective, we studied whether there was a tendency for authorship teams in terms of citation, sentiment score or the frequency of positive/negative terms when citing papers of the same gender composition teams. This objective was answered through questions 2, 3 and 4.

The results of the analysis from question 2 showed a significant tendency for female and male authorship teams to cite papers by the same gender composition teams. This homophily was 14% and 8% of the times for men and women, respectively. This finding makes sense, as found in previous research, networks of men are more homophilic than women's, especially in environments in which men are dominant [70,71]. Our finding is in line with Ghiasi et al. [15], who also found a gender homophily in referencing patterns. The authors concluded that the lower number of female authorships might lead to men's higher inclination (compared with women) to cite papers by the same gender, which in turn could contribute to the Matilda effect in science [12].

The results of regression analysis from question 3 showed a small significant positive association between the gender similarity of cited and citing authorship teams and the sentiment score received, while controlling for several covariates. In other words, as the cosine gender similarity between cited and citing papers increased, the positive sentiment score slightly increased. This finding seems to suggest that there is a slight tendency for authors to be more positive towards authorship teams with a more similar gender composition. However, further analysis (question 4) showed that the frequency of positive terms for male authorship teams was significantly higher than expected when citing a male authorship team. In contrast, the frequency of negative terms for female authorship teams was significantly higher than expected when citing a female authorship team. The findings about the use of positive and negative terms when citing same gender authorship teams is in line with Lerchenmueller et al.'s [3] study which showed that in medicine and life sciences, female first-last authored articles were 21% less probably to present their research work positively compared with articles in which the first and/or last author was male. The finding is also in line with McKinnon and O'Connell's [11] study, which found that in STEM, women academics who publicly communicated their scientific work were probably to be negatively stereotyped, judged and held back by other women. The authors suggested the reason why women may judge other women more harshly, could be simply due to underlying implicit bias. As language statistics have been shown to predict people's implicit gender biases [7], it could also be the case in this study that authors' implicit biases about gender could be translated into sentiments regarding the articles they cite. Another explanation to this finding could be related to men's higher tendency to use more boastful or authoritative language compared with women, who may also use less compelling language and be less probably to self-promote their research findings [4]. Consequently, men could be more positive regarding the work of other men, whereas women may be less positive to the work of their own gender.

When conducting regression analyses for questions 1 and 3, we controlled for several important factors that were associated with sentiment score according to previous research. Among these factors, lay summary for a cited paper, country similarity and publication venue of a cited paper (when it was a mega journal) had a positive significant association with the sentiment score received. The finding about lay summary as also indicated in Tan et al. [53] shows the importance of making the text readable (by providing a lay summary in our case) as it provides a remedy in debiasing the language sentiment effect. The finding regarding mega journal might be related to the Matthew effect of journal's ranking – the tendency for a high-ranked journal to receive more citations to its papers [72]. In addition, our findings showed that authors affiliated to an institute in a non-English-speaking country had a lower positive sentiment when they were citing or being cited. As the use of English language in scientific papers can affect one's judgement about the quality of a scientific paper [73], this might presumably create a bias towards authors affiliated to non-English-speaking countries. The finding with regard citing authors is in line with a study by Kourilová [74], which found that non-native English speakers often wrote in an honest, or brutal, blunt manner in their scientific reports.

Collectively, our findings showed gender differences in citation homophily and citation sentiments in the fields of in life science and biomedicine. In general, our findings seem to suggest the existence of Mathew/Matilda effect in the field of life sciences and biomedicine, where there is a higher tendency for male authors to cite papers written by the same gender. Furthermore, when being cited, male-authored papers received a significant higher positive sentiment compared with female-authored papers. Men also used significantly more positive terms when citing papers written by the same gender composition teams. However, this study seems to suggest that not only men contribute to this gender gap in this field, but also women; as our finding showed a lower tendency for women to use positive terms when citing the research findings of papers with the same gender composition. The higher tendency of men to use more positive, boastful or authoritative language compared with women as shown per pervious research [3–5] might act in two ways: men to be more positive towards their own gender and women be less positive towards their own gender. Previous research suggested the extent to which women promote their research accomplishments relative to men as one of the mechanisms that might contribute to current gender gaps in academic medicine [3]. However, it should also be considered that as the promotion of research findings for women might be considered counter stereotypical behaviour [10,75], women might be criticised for this behaviour, even sometimes by other women [4,11]. Thus, it might be the case that implicit gender biases negatively act against women in the way that their research findings are perceived by both men and women. However, further research is required to corroborate this finding.

We believe that our findings are important, as they provide awareness regarding these gender differences in tendency of authors, in terms of citation and language framing of research findings that they cited. Thus, they could serve as a basis for initial actions that can be taken to reduce the gender gap in life sciences and biomedicine.

Finally, this study has some limitations. In this study, we used SentiWordNet for sentiment analysis. Although this dictionary has a large vocabulary coverage and our results demonstrated a good reliability, this analysis could be potentially further improved using a field specific dictionary, which is not currently available [76]. Alternative algorithms such as supervised machine learning methods could also be used in future research.

Declaration of conflicting interests

Funding

ORCID iD

Tahereh Dehdarirad (iD) https://orcid.org/0000-0003-2529-962X

Supplemental material

Supplemental material for this article is available online.

Notes

1. The text surrounding each cited reference within the text is called *citances* [77].
2. No, not, n't, never, neither, nor, none, nobody, nowhere, nothing, cannot, cannot, without, no one and no way.
3. Synonym sets.
4. The R codes were provided by John Skvoretz at a Social Network Analysis course.
5. To obtain the number of citations for first and last authors, we conducted a search in the SciVal Author Lookup API (https://dev.el-sevier.com/documentation/SciValAuthorAPI.wadl) using authors' IDs.
6. A lay summary is a brief summary of research that uses clear, plain language to communicate complex ideas and technical and scientific terms to an audience who does not have prior knowledge about the subject.

References

[1] Girod S, Fassiotto M, Grewal D et al. Reducing implicit gender leadership bias in academic medicine with an educational intervention. *Acad Med* 2016; 91: 1143–1150.
[2] Montalvo JG, Alimonti D, Reiland S et al. Gender stereotype and the scientific career of women: evidence from biomedical research centers. *bioRxiv* 2020. DOI: 10.1101/2020.10.29.360560.

[3]   Lerchenmueller MJ, Sorenson O and Jena AB. Gender differences in how scientists present the importance of their research: observational study. *BMJ* 2019; 367: l6573–l6573.

[4]   Reuben E, Sapienza P and Zingales L. How stereotypes impair women's careers in science. *Proc Natl Acad Sci U S A* 2014; 111: 4403–4408.

[5]   Witteman HO, Hendricks M, Straus S et al. Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *Lancet* 2019; 393: 531–540.

[6]   Franco MC, Rice DB, Schuch HS et al. The impact of gender on scientific writing: an observational study of grant proposals. *J Clin Epidemiol* 2021; 136: 37–43.

[7]   Lewis M and Lupyan G. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat Hum Behav* 2020; 4: 1021–1028.

[8]   Burkinshaw P. Introduction. In: Taylor Y (ed.) *Higher education, leadership and women vice chancellors: fitting into communities of practice of masculinities*. London: Palgrave Macmillan, 2015, pp. 1–17.

[9]   Rudman LA. Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *J Pers Soc Psychol* 1998; 74: 629–645.

[10]   Azoulay P and Lynn F. Self-citation, cumulative advantage, and gender inequality in science. *Sociol Sci* 2020; 7: 152–186.

[11]   McKinnon M and O'Connell C. Perceptions of stereotypes applied to women who publicly communicate their STEM work. *Humanit Soc Sci Commun* 2020; 7: 160.

[12]   Rossiter MW. The Matthew Matilda effect in science. *Soc Stud Sci* 1993; 23: 325–341.

[13]   Dion ML and Mitchell SM. How many citations to women is 'enough'? Estimates of gender representation in political science. *PS Polit Sci Polit* 2020; 53: 107–113.

[14]   Bertolero M, Dworkin J, David S et al. Racial and ethnic imbalance in neuroscience reference lists and intersections with gender. *bioRxiv* 2020. DOI: 10.1101/2020.10.12.336230.

[15]   Ghiasi G, Mongeon P, Sugimoto C et al. Gender homophily in citations. In: *23rd international conference on science and technology indicators* (eds R Costas, T Franssen and A Yegros-Yegros), Leiden, 12–14 September 2018, pp. 1519–1525. Leiden: Centre for Science and Technology Studies (CWTS).

[16]   Skvoretz J. Diversity, integration, and social ties: attraction versus repulsion as drivers of intra- and intergroup relations. *Am J Sociol* 2013; 119: 486–517.

[17]   Athar A. *Sentiment analysis of scientific citations*. Cambridge: Computer Laboratory, University of Cambridge, 2014.

[18]   Xu J, Zhang Y, Wu Y et al. Citation sentiment analysis in clinical trial papers. *AMIA Annu Symp Proc* 2015; 2015: 1334–1341.

[19]   Ghosh S, Das D and Chakraborty T. *Determining sentiment in citation text and analyzing its impact on the proposed ranking index*. Cham: Springer International Publishing, 2018.

[20]   Yan E, Chen Z and Li K. The relationship between journal citation impact and citation sentiment: a study of 32 million citances in PubMed Central. *Quant Sci Stud* 2020; 1: 664–674.

[21]   Zhang G, Ding Y and Milojević S. Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content. *J Am Soc Inf Sci Technol* 2013; 64(7): 1490–1503.

[22]   Kilicoglu H, Peng Z, Tafreshi S et al. Confirm or refute? A comparative study on citation sentiment classification in clinical research publications. *J Biomed Inform* 2019; 91: 103123.

[23]   Athar A and Teufel S. Context-enhanced citation sentiment detection. In: Fosler-Lussier E, Riloff E and Bangalore S (eds) *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies*. Montreal, QC, Canada: Association for Computational Linguistics, 2012, pp. 597–601.

[24]   Sadia A, Khan FK and Bashir F. An overview of lexicon-based approach for sentiment analysis. In: *3rd international electrical engineering conference (IEEC 2018)*, 9–10 February 2018, Karachi, Pakistan.

[25]   Goeuriot L, Na J-C, Min Kyaing WY et al. Sentiment lexicons for health-related opinion mining. In: Yang CC (ed.) *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*. Miami, FL: Association for Computing Machinery, 2012, pp. 219–226.

[26]   Wiley MT, Jin C, Hristidis V et al. Pharmaceutical drugs chatter on Online Social Networks. *J Biomed Inform* 2014; 49: 245–254.

[27]   del Pilar Salas-Zárate M, Alor-Hernández G, García-Alcaraz JL et al. A sentiment analysis method for analyzing users opinions about drugs for chronic diseases. In: Vasant P, Litvinchev I, lejo-Saucedo JAM, et al. (eds) *Data analysis and optimization for engineering and computing problems* 2020: EAI/Springer innovations in communication and computing. Cham: Springer, 2020.

[28]   Yaghtin M, Sotudeh H, Nikseresht A et al. Modeling the co-citation dependence on semantic layers of co-cited documents. *Online Inf Rev.* Epub ahead of print 2021. DOI: 10.1108/OIR-04-2020-0126.

[29]   Cavalcanti DC, Prudêncio RBC, Pradhan SS et al. Good to be bad? Distinguishing between positive and negative citations in scientific impact. In: *2011 IEEE 23rd international conference on tools with artificial intelligence*, Boca Raton, FL, 7–9 November 2011, pp. 156–162. New York: IEEE.

[30]   Sendhilkumar S, Elakkiya E and GS M. Citation semantic based approaches to identify article quality. In: *The third international conference on computer science, engineering and applications-ICCSEA* (ed DC Wyld), Delhi, India, 24–26 May 2013, pp. 411–420. Cham: Springer.

[31] Gonçalves P, Araújo M, Benevenuto F et al. Comparing and combining sentiment analysis methods. In: Abbadi AE and Krishnamurthy B (Eds) *Proceedings of the first ACM conference on online social networks*. Boston, MA: Association for Computing Machinery, 2013, pp. 27–38.

[32] Jia M. *Citation function and polarity classification in biomedical papers*. London, ON, Canada: University of Western Ontario, 2018.

[33] Agarwal S, Choubey L and Yu H. Automatically classifying the role of citations in biomedical articles. *AMIA Annu Symp Proc* 2010; 2010: 11–15.

[34] PubMed Central. Open access subset, 2021, https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[35] Fujiwara T and Yamamoto Y. Colil: a database and search service for citation contexts in the life sciences domain. *J Biomed Semant* 2015; 6: 38.

[36] Esuli A and Sebastiani F. *SentiWordNet: a high-coverage lexical resource for opinion mining*. Technical report ISTI-PP-002/2007. Rome: Institute of Information Science and Technologies of the Italian National Research Council, 2007.

[37] Mahalakshmi GS, Siva R and Sendhilkumar S. *Context based retrieval of scientific publications via reader lens*. New Delhi, India: Springer, 2015.

[38] Denecke K. Sentiment analysis from medical texts. In: Denecke K (ed.) *Health web science: social media data for healthcare*. Cham: Springer, 2015, p. 87.

[39] Asghar MZ, Ahmad S, Qasim M et al. SentiHealth: creating health-related sentiment lexicon using hybrid approach. *Springerplus* 2016; 5: 1139.

[40] Yaghtin M, Sotudeh H, Mirzabeigi M et al. In quest of new document relations: evaluating co-opinion relations between co-citations and its impact on Information retrieval effectiveness. *Scientometrics* 2019; 119: 987–1008.

[41] Rotondi MA and Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. *J Clin Epidemiol* 2012; 65: 778–784.

[42] Gamer M, Lemon J and Fellows Puspendra SI. irr: various coefficients of interrater reliability and agreement, 2019, https://cran.r-project.org/package=irr

[43] McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* (Zagreb) 2012; 22: 276–282.

[44] Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.

[45] Bolker BM. Linear and generalized linear mixed models. In: Gordon SN-Y, Fox A and Sosa VJ (eds) *Ecological statistics: contemporary theory and application*. Oxford: Oxford University Press, 2015, p. 315.

[46] Golub MS and Sobin CA. Statistical modeling with litter as a random effect in mixed models to manage 'intralitter likeness'. *Neurotoxicol Teratol* 2020; 77: 106841.

[47] West BT, Welch KB and Gałecki AT. *LINEAR MIXED MODELS: A Practical Guide Using Statistical Software*. Ann Arbor, USA: CRC Press, Taylor & Francis Group, 2015.

[48] Ralitza G. Linear mixed models for longitudinal and clustered data. In: Ralitza G (ed.) *Statistical methods in psychiatry and related fields: longitudinal, clustered, and other repeated measures data*. Boca Raton, FL: CRC Press Taylor & Francis Group, 2017.

[49] Pinheiro J, Bates D, DebRoy S et al. nlme: linear and nonlinear mixed effects models (R package version 3.1-152), 2021, https://CRAN.R-project.org/package=nlme

[50] Lenth RV, Buerkner P, Herve M et al. Emmeans: estimated marginal means, aka least-squares means, 2021, https://cran.r-project.org/web/packages/emmeans/index.html

[51] Lüdecke D. sjPlot: data visualization for statistics in social science (R package version 2.8.9), 2021, https://CRAN.R-project.org/package=sjPlot

[52] Pancer E, Chandler V, Poole M et al. How readability shapes social media engagement. *J Consum Psychol* 2019; 29: 262–270.

[53] Tan H-T, Ying Wang E and Zhou B. When the use of positive language backfires: the joint effect of tone, readability, and investor sophistication on earnings judgments. *J Account Res* 2014; 52: 273–302.

[54] Petersen AM, Fortunato S, Pan RK et al. Reputation and impact in academic careers. *Proc Natl Acad Sci U S A* 2014; 111: 15316–15321.

[55] Vinkers CH, Tijdink JK and Otte WM. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *BMJ* 2015; 351: h6467.

[56] Santamaría L and Mihaljević H. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput Sci* 2018; 4: e156.

[57] International Committee of Medical Journal Editors (ICMJE). Uniform requirements for manuscripts submitted to biomedical journals: defining the role of authors and contributors, 2010, http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

[58] Sharma H and Verma S. Authorship in biomedical research: a sweet fruit of inspiration or a bitter fruit of trade. *Trop Parasitol* 2018; 8: 62–69.

[59] Chapman B and Chang J. Biopython: python tools for computational biology. *SIGBIO Newsl* 2000; 20: 15–19.

[60] Leydesdorff L, Comins JA, Sorensen AA et al. Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level. *Scientometrics* 2016; 109(3): 2077–2091.

[61] Shu F, Qiu J and Larivière V. Mapping the biomedical sciences using Medical Subject Headings: a comparison between MeSH co-assignments and MeSH citation pairs. *J Med Libr Assoc* 2021; 109(3): 441–449.

[62] Mishra S, Fegley BD, Diesner J et al. Self-citation is the hallmark of productive authors, of any gender. *PLoS ONE* 2018; 13: e0195773.

[63] Dehdarirad T. Could early tweet counts predict later citation counts? A gender study in Life Sciences and Biomedicine (2014–2016). *PLoS ONE* 2020; 15: e0241723.

[64] Andersen JP, Schneider JW, Jagsi R et al. Gender variations in citation distributions in medicine are very small and due to self-citation and journal prestige. *eLife* 2019; 8: e45374.

[65] Spezi V, Wakeling S, Pinfield S et al. Open-access mega-journals. *J Document* 2017; 73: 263–283.

[66] Shailes S. Something for everyone. *eLife* 2017; 6: e25411.

[67] Didegah F, Bowman TD and Holmberg K. On the differences between citations and altmetrics: an investigation of factors driving altmetrics versus citations for Finnish articles. *J Assoc Inform Sci Technol* 2018; 69: 832–843.

[68] Benoit K, Watanabe K, Wang H et al. quanteda: an R package for the quantitative analysis of textual data. *J Open Source Softw* 2018; 3: 774.

[69] Kleiman E. Understanding and analyzing multilevel data from real-time monitoring studies: an easily accessible tutorial using R, 2017, osf.io/9qv7n

[70] Brass DJ. Men's and women's networks: a study of interaction patterns and influence in an organization. *Acad Manag J* 1985; 28: 327–343.

[71] Ibarra H. Homophily and differential returns: sex differences in network structure and access in an advertising firm. *Adm Sci Q* 1992; 37: 422–447.

[72] Drivas K and Kremmydas D. The Matthew effect of a journal's ranking. *Res Policy* 2020; 49(4): 103951.

[73] Eve MP, Neylon C, O'Donnell DP et al. *Reading peer review: PLOS ONE and institutional change in academia. Elements in publishing and book culture.* Cambridge: Cambridge University Press, 2021.

[74] Kourilová M. Communicative characteristics of reviews of scientific papers written by non-native users of English. *Endocr Regul* 1998; 32: 107–114.

[75] Pritlove C, Juando-Prats C, Ala-Leppilampi K et al. The good, the bad, and the ugly of implicit bias. *Lancet* 2019; 393: 502–504.

[76] Denecke K and Deng Y. Sentiment analysis in medical settings: new opportunities and challenges. *Artif Intell Med* 2015; 64: 17–27.

[77] Nakov P, Schwartz A and Hearst M. Citances: citation sentences for semantic analysis of bioscience text. In: *SIGIR '04 workshop on search and discovery in bioinformatics*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.2666&rep=rep1&type=pdf