Scientometrics (2019) 120:1225–1235
https://doi.org/10.1007/s11192-019-03160-6

The language of peer review reports on articles published in the BMJ, 2014–2017: an observational study

Alberto Falk Delgado[1] · Gregory Garretson[2] · Anna Falk Delgado[3]

Received: 9 February 2019 / Published online: 28 June 2019

Abstract

To analyse the words and expressions used in peer reviews of manuscripts that were later published as original research in the BMJ. Secondary aims were to estimate the differences in net sentiment between peer review reports on manuscripts subject to one or more rounds of peer review and and review reports on initially rejected manuscripts that were accepted after appeal. This observational study included all peer review reports published in the BMJ from September 2014 until the end of 2017. The study analysed the frequency of specific words in peer review reports for accepted manuscripts, identifying the most commonly occurring positive and negative words and their context, as well as the most common expressions. It also quantified differences in net sentiment in peer review reports between manuscripts accepted after appeal and manuscript accepted without appeal. The dataset consisting of 1716 peer review reports contained 908,932 word tokens. Among the most frequent positive words were “well”, “important”, “clear”, “while the negative words included “risk”, “bias”, and “confounding”. The areas where the reviewer makes the most positive and negative comments included: “well-written paper”, “well-written manuscript”, “this is an important topic”, “answers an important question”, “high risk of bias” and “selection bias”. The sentiment analysis revealed that manuscripts accepted after appeal had lower scores on review reports for joy and positive sentiment, in addition to having higher scores for negative words expressing sadness, fear, disgust and anger compared with manuscripts that were not initially rejected. Peer review comments were mainly related to methodology rather than the actual results. Peer review reports on initially rejected manuscripts were more negative and more often included expressions related to a high risk of bias.

Keywords BMJ · Peer review · Reviewer comments · Linguistics · Medicine · Open peer review

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11192-019-03160-6) contains supplementary material, which is available to authorized users.

✉ Alberto Falk Delgado
Alberto.falk-delgado@surgsci.uu.se

Extended author information available on the last page of the article

Introduction

In the academic publishing process, the peer review system plays an essential role, with the aim and potential to significantly contribute to the quality and rigor of the work published (Fletcher and Fletcher 2003). Most researchers have experienced both positive and, perhaps more commonly, negative peer reviews. In a previous and unpublished study, peer review reports were collected from a single-blind randomized controlled trial (RCT); the aim of the RCT was to assess the impact of training on reviewers in order to improve peer review reports on three manuscripts. The peer review reports collected from the RCT were analysed using computational linguistic software. The study, comprising 440 individual peer review reports, suggested that the peer review reports on manuscripts recommended for rejection described the manuscript as more "authentic" (such as more honest and more disclosive) than manuscripts that were recommended for publication by the reviewers (Glonti et al. 2017). In this previous report, 330 reviews recommended rejection, so little is known about the vocabulary used in peer review reports on manuscripts that are accepted. One obvious problem in conducting research on authentic peer reviews has been that the peer review reports have not, until recently, been publicly available.

Although a review can include a recommendation to reject or accept the manuscript, the decision is ultimately editorial. The central role of the editor in the decision to reject or accept a manuscript is evident in the form of review reports used by the BMJ, where a recommendation to reject or accept a manuscript is *not* requested from reviewers today, nor is a reviewer requested to provide a score for the manuscript, which is different from most of the other medical journals where a recommendation is requested from the referee'. Instead, the journal asks a reviewer to assess the the manuscript according to a number of criteria such as originality, importance, relevance to a general medical audience, and scientific robustness, with descriptions in textual form and not using ordinal categories. Since 2014, the BMJ has had a fully open peer review, meaning that all research articles have their prepublication history made available on the homepage alongside a fully signed report from each reviewer (The BMJ). It is though, according to the editors, that signed reviews may be more constructive than anonymous reviews, (McNutt et al. 1990) and provide greater accountability for peer review reports (Groves 2010; Groves and Loder 2014). In addition, as a positive side effect, the open peer review system gives credit to the reviewers.

The aim of the present study was to analyse the words and expressions used in peer reviews when reporting on manuscripts that were later published as original research in the BMJ. A secondary aim was to estimate the differences in net sentiment between peer review reports of manuscripts subject to one or more rounds of peer review and review reports on manuscripts initially rejected but accepted after appeal.

Methods

Study design

This observational mixed-design study included all peer review reports published in the BMJ from September 2014 (inception of open peer review policy) until the end of 2017.

Selection criteria

Only peer review reports on original research manuscripts were eligible for inclusion. Reviews and other non-original research articles were excluded. The data included reports on accepted manuscripts and manuscripts initially rejected but accepted after appeal, but did not include rejected manuscripts, since their peer reviews are not published.

Data sources and data collection

Data for the study was obtained from the BMJ website (www.bmj.com). All studies from the beginning of September 2014 were assessed for inclusion—that is, from the inception of open peer review. Studies published before August 2017 were included retrospectively; studies after that date were included as they were released. For each included research article, the first author's name and the publication date were retrieved. The number of peer review rounds was registered, as well as the number of peer review reports per peer review round. Each peer review report was saved as a text file, excluding the identity of the peer reviewer. The discussions from the editorial meeting were excluded, since they are of a more general character, and not strictly part of the peer review report. Each individual peer review report was saved to a unique file, providing a single-peer-review resolution of the data and enabling comparison between peer review rounds. The usage of specific words could thus be traced to the original article, providing a context for the usage. If peer review reports were missing from the second round, the peer reviews could also be included from the authors' responses when they were not available elsewhere. In cases of appeal, when a manuscript was rejected, the rejection decision along with the review reports were defined as round one, the appeal as round two, and any later peer reviews as round three.

Outcome measures

The main outcome measure was the frequency of specific words (excluding function words such as "and", "that", and "the") in the peer review reports for accepted manuscripts in the BMJ. Secondary outcomes included:

(a) Identifying the most commonly occurring positive and negative words, and their context of usage.
(b) Comparing the most common expressions (sequences of 5 words occurring together) in peer review reports on initially accepted manuscripts versus manuscripts that were accepted after appeal.
(c) Quantifying the differences in net sentiment in between reports on manuscripts accepted after appeal and manuscript accepted without appeal.

Statistical analysis

To count the frequency of all words occurring in peer review reports, the program AntConc was used (Anthony L, version 3.4.4, Waseda, Japan, For Mac). All separate text files were merged into one single text file, and then stratified for: the first round of peer review reports on accepted manuscripts (both reports on manuscripts that had one revision and reports on manuscripts that had more than one revision), and the first round of peer review reports on manuscripts that were initially rejected but accepted after appeal (see Fig. 1). the text files were merged using the cat command function in Mac OS (10.10.5). The preprocessing of the texts was performed by removing stop words (the most common short function words in a language such as “the”, “has”, “have” etc.) derived from three lexicons, provided by “tidy-text” (version 0.1.7), a package for text mining in R (version 10.10.5). We then identified words occurring at least 1000 times in all peer review reports. Further, we identified the most common adjectives using the program AntConc (version 3.4.4). In order to identify the most common positive and negative words in the peer review reports, we used in a sentiment lexicon (Liu 2018). For the two most common positive words, the five most frequently associated words were evaluated in bigrams e.g. “well-written”, “well-done”, “important topic”, “important question” etc. This was performed also for the two most commonly occurring negative words e.g. “high risk”, “increased risk”, “publication bias”, and “selection bias”. We normalized for hypehenation, counting e.g. “well-known” and “well known” as equivalent.

Naturally, some words can be used in a negative linguistic context (compare “satisfactory” and “not satisfactory”). To test the relationship between the word “not” and other words, we used the bigram function in the R package “dplyr”, since a positive word preceded by “not” might yield misleading in the single-word analysis. Other similar constructions were also tested e.g. insufficient, lacking.

Further, we identified the most common sequences of words. We chose to identify the five most frequent 5-grams by using the program AntConc (version 3.4.4). We chose five words, since shorter strings were less context-specific, and longer strings are more infrequent. No ‘stop words’ were removed for this purpose.

In the sentiment analysis three sentiment lexicons were used namely Bing (Liu 2018), AFINN (Nielsen 2011) and NRC(Turney 2013), which define words as either positive or negative. The AFINN lexicon contains 2477 words, with each word having a score between −5 (highly negative) and 5 (highly positive). The Bing lexicon contains 6788 words, each simply defined as either positive or negative. The NRC lexicon contains 14,182 words, with each word classified as representing one of eight emotions—anger, anticipation,

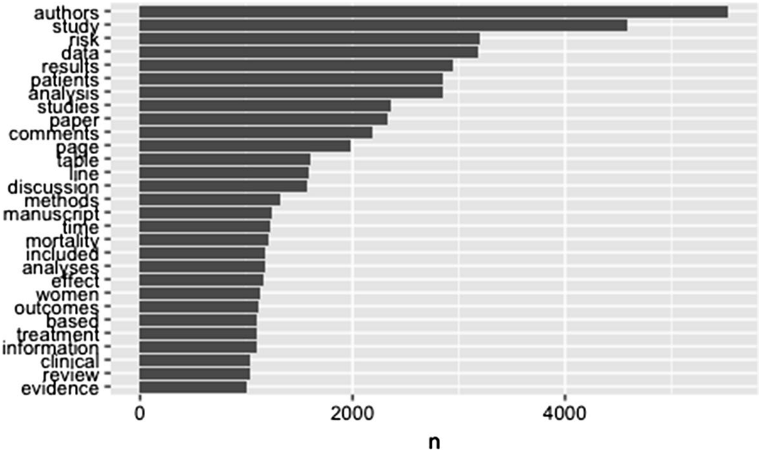

Fig. 1 The most common words occurring at least 1000 times in all peer review reports after removal of stop words

disgust, fear, joy, sadness, surprise, or trust—and two sentiments—positive or negative. All lexicons have more negative words than positive words, but this disparity is especially great in the Bing lexicon.

To quantify the difference in sentiment and emotion between types of peer review reports, we conducted pairwise comparisons in net sentiment between manuscripts accepted after appeal and manuscripts accepted without appeal. The difference in sentiment and emotion was calculated using a Poisson test and plotted with 95% confidence intervals. All plots were made using “ggplot2” package in R.

Results

After screening all original manuscripts published between September 2014 and the end of 2017, all published original studies ($n = 520$; Supplementary Table 1) were further assessed for the presence of an associated published review report. Out of 520 published original articles, 365 had at least one associated published review report. One hundred fifty-five manuscripts lacked an open peer review report; this was more common shortly after the inception of open peer reviews in the BMJ in the years 2014 to 2015. The 365 manuscripts had a total of 1716 published single-peer review reports, with a median of 3 (range 2–4) peer review reports for each manuscript. Two hundred twenty-seven of the studies had only one round of peer review report, and 138 studies had more than one round of peer review, while only two studies had five rounds of peer review reports. Nineteen studies had initially been rejected but were later accepted after the authors had written to appeal the decision.

The language of peer review reports

The dataset consisting of 1716 peer review reports contained 908,932 word tokens (individual instances), representing 19,177 different types (unique words). The most common word was “the”, appearing 61,589 times, followed by “of”, with a frequency of 30,190, and “to”, with a frequency of 22,327. Removal of such function words, and all words occurring fewer than 1000 times, produced a set of words mainly related to the methods and results of the study, such as “risk”, “data”, “analyses”, and “outcomes”; see Fig. 1. Among adjectives/adverbs, the most common words were (frequency): “well” (1759), “important” (1731), “very” (1636), “included” (1232), “different” (1189), “clear” (1125), “first” (1004), “many” (976), “high” (926), “general” (814), and “relevant” (805).

Sentiment analysis

The 30 most frequent positive words in all peer review reports included “well”, “important”, “clear”, “significant”, “strong”, and “interesting”. The 30 most common negative words included “risk”, “bias”, “confounding”, “difficult”, “unclear”, “concern” and “problem”, as seen in Fig. 2. The word “well” occurred (frequency) in the expressions “well-written” (332), “well-done” (71), “well-known” (63), “well-conducted” (62), and “well-described” (59). “Important” was associated with “topic” (66), “question” (47), “information” (35), “contribution” (34), and “issue” (34). “Risk” was associated with “high” (224), “increased” (216), “low” (100), “relative” (63), and “absolute” (57). “Bias” was associated with “publication” (76), “selection” (65), “potential” (30), “time” (16) and “reporting” (12).

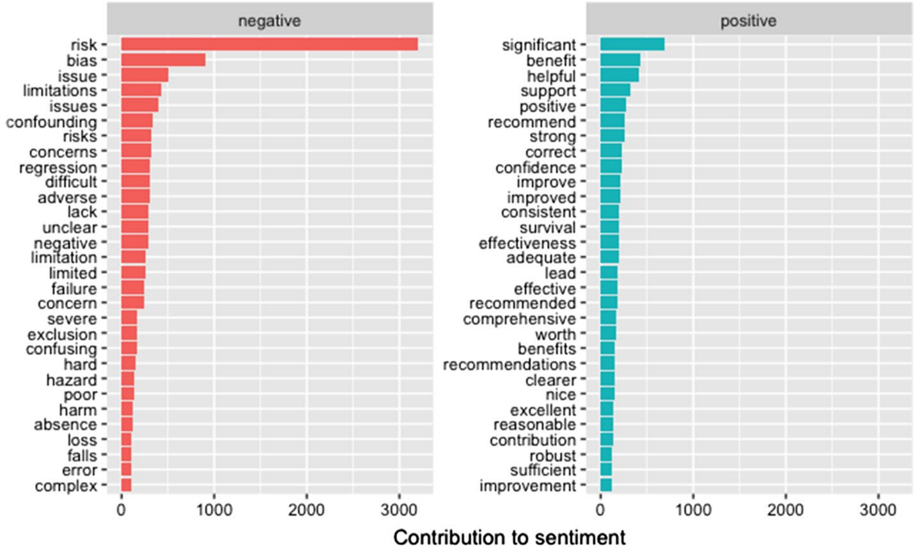

Fig. 2 The 30 most common positive and negative words occurring in peer review reports

We then identified the areas in which the reviewer made the most positive and negative comments: "well-written paper" (15), "well-written manuscript" (8), "well-written article" (6), "well-written study" (6), "well done study" (4), "well-conducted study" (14). The phrase "important topic" occurs in contexts such as "this is an important topic" (10), "paper on an important topic" (3), and "study on an important topic" (3), while "important question" occurs in "is an important question" (8), "addresses an important question" (5), "answer an important question" (4). The areas related to risk were: "high risk of bias" (48)—mainly relating to the study design—and "associated with increased risk" (16), relating mainly to the results. Issues relating to publication bias were commonly discussed in peer review reports on meta-analyses, while "selection bias" (65) is mainly discussed in terms of the study design, e.g. "The study cohort can nevertheless be considered as well protected against selection bias". In contrast, there appears to be little mention of the actual results of the study: "interesting findings" (6), "good results" (2), "interesting results" (2), "strong results" (2).

Since a positive word might be preceded by "not", a single positive word can easily be misleading. We identified all words followed by "not"; for example, "not" was followed by "clear" 330 times, "clearly" 38 times, and "convinced" 36 times; see Fig. 3. Other ways of negating a concept were also investigated. "Insufficient" was preceded or followed by "is" (11), "evidence" (7), "data" (5), "in" (4). "Lacking" was preceded or followed by "is" (9), "in" (5), "be" (4), and "are" (3). We further assessed ways of hedging a negative assessment by using a positive term such as "clear". "Would be clearer" (18), "be made clearer" (14), "should be clearer" (9), and "could be clearer" (9), were all found to occur. Further, we investigated the "not…as" construction, and found "not as clear" (4), "not as good" (3), and "not as safe" (3).

Finally, we conducted a net sentiment analysis to determine whether there were differences between peer review reports on manuscripts that underwent one or several rounds of revision on one hand, versus peer review reports on initially rejected manuscripts accepted

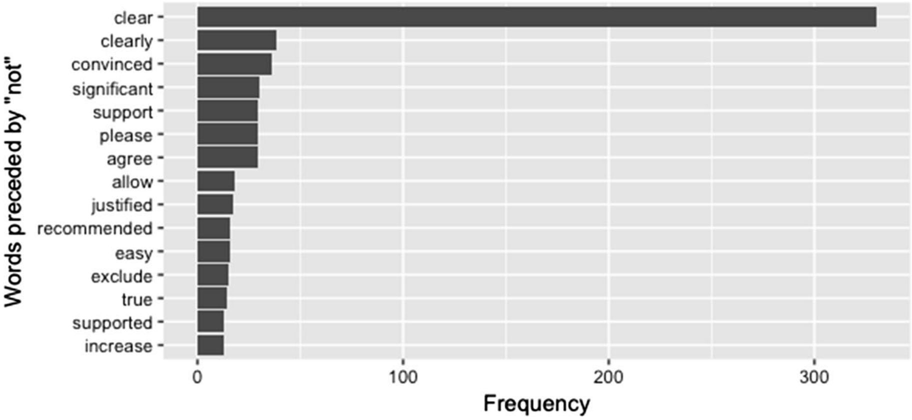

Fig. 3 The 20 most common sentiment words preceded by “not” (note that these can be either positive or negative)

after appeal on the other hand. In all cases, only the first round of reports was considered. The sentiment analysis revealed that reports on manuscripts accepted after appeal had lower scores for joy and positive sentiment, in addition to having higher scores for negative emotions such as sadness, fear, disgust and anger; see Fig. 4.

N-gram analysis

We investigated the most common five-word n-grams (sequences of words) to identify patterns in the reviews. A common reviewer statement mentioned or suggested an improvement, as in “it would be helpful/useful/interesting to”, for manuscripts that underwent one or several rounds of revision; see Supplementary Table 2. Another common expression was “clearly defined and appropriately answered”. In addition “importance of work to

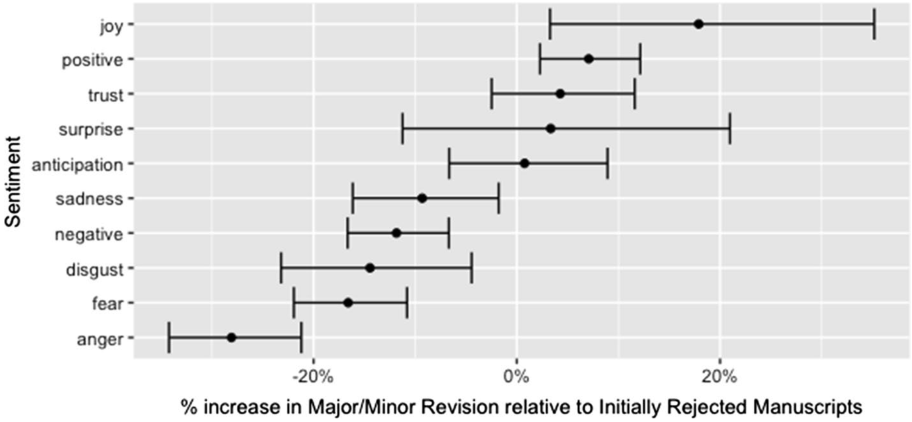

Fig. 4 Percentage increase in sentiment in one or more rounds of revision (major/minor revision) relative to initially rejected manuscripts that were accepted after appeal

general/of work to general readers" were among the most common expressions for manuscripts with one or more revisions. For manuscripts initially rejected and later accepted after appeal, concerns related to biases were common, such as "at high risk of bias" and "too high risk of bias/a high potential for bias/high risk of bias due/overall high risk of bias". Claims that the manuscript was well written were also common in this group: "the paper is well written/this is a well-written".

Discussion

Principal findings

This study, based on 1716 peer review reports on 365 unique manuscripts accepted by the BMJ, revealed that positive words primarily concerned the manuscript's research question, the methodological rigor and the quality of the writing. Manuscripts initially rejected but later accepted were commonly described in peer review reports as being "at high risk of bias". Words describing the article's limitations were common, such as "bias", "confounding" and "unclear". Peer reviews of studies initially rejected by the BMJ include significantly more negative words and fewer positive words than manuscripts that were not initially rejected.

Strengths of the study

The main strength of this study is related to the large sample size used, including all publicly available peer review reports for all manuscripts accepted by the BMJ between 2014 and 2017. In order to minimize any bias data were analysed using non-arbitrary and reproducible methods, not depending on manual classification, which would depend on subjective individual judgment.

Limitations of the study

The limitations of the study relate to the fact that the manuscripts were published in only one journal—the BMJ. Since the BMJ has a general readership, the generalizability might be limited to other such journals. The BMJ provides a focused template for their referees related to methodology, importance of the research question and how well the manuscript is written, therefore it is unclear if the findings in this study can be generalizable up to journals with other peer review templates. It is possible that evidence based reporting guidelines for transparent and reproducible research such as STROBE, CONSORT and PRISMA depending on research design might have influenced the language used in referee reports. Furthermore, because we only included studies that were (eventually) accepted for publication, peer review reports leading to manuscript rejection could not be included in the study, leading to a selection bias given the large difference in size of the analysed groups. A further limitation is that we have not been able to study how the results might have been influenced by the fact of the peer review process being open. Closed peer reviews are by definition, not available for comparison. In addition, a yet further limitation of the study is the fact that peer review reports were anonymized and that this might have lead to another potential bias, we don't know if there were repeat reviewers in the analysed group and whether this influenced the results, as a reviewer might be prone to using the same or similar words and expressions in their review reports. Yet another potential bias of the study is that there were some missing reports from the authors' responses and that these reports might not have been cited in their entirety. The actual scoring of a word as positive or negative is based on the combination of three lexicons, however they have not been validated specifically for this purpose.

Strengths and limitations of the study in relation to other studies

Current knowledge of the language used in the peer reviews of accepted manuscripts is limited. In the conference paper by Ketevan Glonti mentioned above, based on 440 peer review reports, in which each peer review report also included a recommendation on whether the manuscript should be accepted or not, the language in the peer review reports was classified into four categories (Glonti et al. 2017). However, the reports analysed in the Ketavan Glonti study were all based on three manuscripts, and in contrast to current practice, these reports included recommendations for acceptance or rejection. In addition, these peer review reports were of a hypothetical nature and not actually used for editorial decisions, possibly limiting the external validity of the findings. Another study by Bornmann, Lbased on peer review reports on manuscripts rejected after submission to a top chemistry journal and later accepted in a "low impact" journal or a "high impact" journal, suggested that publication in the latter is unlikely if the peer reviews are negative with reference to novelty or the study design, which also is in line with the results of our study (Bornmann et al. 2009) In a study by Philippa Mungra et al. (2010) reviewer comments from submitted papers (in total 17 manuscripts) to medical journals were gathered among researchers in an medical school, comments were divided broadly into content and language use comments. Content comments (56%) regarding incomplete literature occurred 9.84%, and lack of association between claim and data was found in 9.29%. Among language use comments (44%) comments related to not well written English 7.92% and lack of clarity in 6.83%. This is in line with our data suggesting that most reviewer comments are related to scientific and methodological content. In the study by Fortanet, 50 referee reports were collected from the fields business organisation and applied linguistics (Fortanet 2008). Fortanet found that referees preferred to present criticism followed by recommendations to the manuscript.

Implication of findings

Among the most common positive words used by the reviewers were "well", "strong", "important" and "clear". Positive words were commonly associated with words describing the strengths of the manuscript, such as "well-written", "well-done", "important issue", "important question", etc. The areas which the peer reviewer most commonly commented on were the research question and the methodology, such as study design and risk of bias in the study. The peer review reports commonly discussed how well the text was written, but also how well the research question was answered. However, a bit surprisingly, very few comments were made on the actual results/finding of the study, which suggests that the study design and the research question are more important than the actual findings of the study to reviewers.

Our analysis of the most frequent five-word sequences revealed that "relevance to a general readership" was common, but also that "the research question was clear/was clear and appropriately answered", as well as statements that the manuscript was "well-written". In review reports of initially rejected manuscripts, expressions related to a “high risk of bias” were common. Quantifying the difference in net sentiment showed that reviews of manuscripts with an editorial decision leading to one or several rounds of revision (major/minor revision) were more positive than those for initially rejected manuscripts.

To summarize our findings, it seems unlikely that a manuscript will survive peer review and be accepted in the BMJ unless the peer review reports are associated with positive language stating that the manuscript is well-written, that the methodology is solid, and that the content is appropriate for a wide readership.

This report has aimed to help de-mystify the peer review process and emphasize the focus of this process on the actual scientific content of the work presented. Making researchers aware of the wording of peer reviews and the subsequent editorial decisions can potentially help to level the playing field when it comes to the evaluation of submissions to the BMJ.

Unanswered questions and future research

It is unclear whether the language of peer review reports in the BMJ is generalizable outside the medical field and/or to more specialized journals. Further, it is still unclear how the results would vary with the blinding of peer reviewers and editors. Finally, it is unclear whether manuscripts that receive fewer positive and more negative reviews could still be accepted in large open-access journals. Finally, It would be of interest to compare the language of peer review report template and reporting guidelines with the language of peer review reports in a future analysis.

Conclusions

This study revealed that positive words primarily concerned the manuscript’s research question, the methodological rigour and the quality of the writing featured prominently in reviewers recommendations that a submitted article be accepted in the BMJ. Manuscripts initially rejected but later accepted were commonly described in peer review reports as being “at high risk of bias”. Words describing the article’s limitations were common, such as “bias”, “confounding” and “unclear”. Peer review reports on initially rejected manuscripts were more negative and more often included expressions related to a high risk of bias.

Author contributions AnFD and AlFD has extracted data and conducted the analyses. AnFD, AlFD and GG have analysed the results. AnFD, AlFD and GG have been involved in the drafting of the manuscript. AlFD is guarantor of the article.

Compliance with ethical standards

Conflict of interest The author declares that they have no conflict of interest.

References

Bornmann, L., Weymuth, C., & Daniel, H.-D. (2009). A content analysis of referees' comments: How do comments on manuscripts rejected by a high-impact journal and later published in either a low or high-impact journal differ? *Scientometrics, 83,* 493–506. https://doi.org/10.1007/s11192-009-0011-4.

Fletcher, R. H., & Fletcher, S. (2003). *The effectiveness of editorial peer review*. London: BMJ Books.

Fortanet, I. (2008). Evaluative language in peer review referee reports. *Journal of English for Academic Purposes, 7,* 27–37.

Glonti, K., Hren, D., Carter, S., & Schroter, S. (2017). Linguistic features in peer reviewer reports: How peer reviewers communicate their recommendations. https://peerreviewcongress.org/prc17-0234.

Groves, T. (2010). Is open peer review the fairest system? Yes. *BMJ, 341,* c6424. https://doi.org/10.1136/bmj.c6424.

Groves, T., & Loder, E. (2014). Prepublication histories and open peer review at the BMJ. *BMJ, 349,* g5394. https://doi.org/10.1136/bmj.g5394.

Liu, B. (2018). A list of English positive and negative opinion words or sentiment words https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. Accessed 25 Feb 2018.

McNutt, R. A., Evans, A. T., Fletcher, R. H., & Fletcher, S. W. (1990). The effects of blinding on the quality of peer review: A randomized trial. *JAMA, 263*(10), 1371–1376.

Nielsen, F. Å. (2011). Evaluation of a word list for sentiment analysis in microblog. In *Big things come in small packages 718 in CEUR workshop proceedings* (pp. 93–98).

Philippa Mungra, P. W. (2010). Peer review process in medical research publication. *English for Specific Purposes, 29*(1), 43–53. https://doi.org/10.1016/j.esp.2009.07.002.

The BMJ. (2018). http://www.bmj.com/about-bmj/publishing-model. Accessed 03 May 2018.

Turney, S. M. A. P. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

Affiliations

Alberto Falk Delgado[1] · **Gregory Garretson**[2] · **Anna Falk Delgado**[3]

[1] Plastic and Reconstructive Surgery, Department of Surgical Sciences, Uppsala University, Ing. 78/79 Plastikmottagningen, Akademiska Sjukhuset, 75185 Uppsala, Sweden

[2] Department of English, Uppsala University, Uppsala, Sweden

[3] Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden