



CHALMERS
UNIVERSITY OF TECHNOLOGY

Use of positive terms and certainty language in retracted and non-retracted articles: The case of biochemistry

Downloaded from: <https://research.chalmers.se>, 2025-01-16 11:58 UTC

Citation for the original published paper (version of record):

Dehdarirad, T., Schirone, M. (2023). Use of positive terms and certainty language in retracted and non-retracted articles: The case of biochemistry. *Journal of Information Science*, In Press.
<http://dx.doi.org/10.1177/01655515231176650>

N.B. When citing this work, cite the original published paper.

Use of positive terms and certainty language in retracted and non-retracted articles: The case of biochemistry

Journal of Information Science

1–11

© The Author(s) 2023

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/01655515231176650

journals.sagepub.com/home/jis**Tahereh Dehdarirad** 

Department of Communication and Learning in Science, Chalmers University of Technology, Sweden

Marco Schirone 

Department of Communication and Learning in Science, Chalmers University of Technology, Sweden; The Swedish School of Library and Information Science, University of Borås, Sweden

Abstract

This study aimed to compare retracted (due to misconduct) and non-retracted articles in biochemistry, in terms of proportion of positive terms, certainty score and different certainty aspects. The data set of this study composed of 662 retracted and non-retracted articles published in the time period of 2018–2020 and indexed in Scopus. These 662 articles accounted for 331 non-retracted and 331 retracted articles, which were matched using matching and covariate balancing analysis. The analysis in this article was done using several regression models. Regarding the use of positive terms, the findings showed that retracted articles were 16% less probable to use positive terms in abstracts, titles and findings presented in conclusion and discussion compared with non-retracted articles. In addition, the results regarding the analysis of certainty language, showed that retracted articles were 15% less probable to use certain language, measured by certainty score, in presenting their scientific findings. Finally, regarding the certainty aspects, the results of regression models showed that retracted articles had 11% less likelihood to present their research findings using certain probability aspect.

Keywords

Biochemistry; certainty language; language analysis; matching and covariate balancing; positive terms; retraction; scientific misconduct

1. Introduction

Research misconduct is an important ethical concern affecting the integrity of the biomedical literature [1,2]. Retractions are made for various reasons. Misconduct has been reported to account for the majority of retractions in many disciplines [3]. Retractions are a key proxy for recognising errors in research and for reconciling misconduct in scientific literature. Thus, understanding the characteristics associated with retractions can provide insight and guide policy for journal editors and authors within a discipline [2]. Characteristics of retracted articles have been studied in different biomedical and medical disciplines, such as cancer [4], surgery [5], emergency medicine [6], nursing [7] and veterinary medicine and animal health [2]. The most common characteristics that have been studied in several research works were journal impact factor [8], affiliation country or institution [2,9], reasons for retraction [2,5], accessibility of data regarding retraction notices [10] and publisher [9]. Very few studies have been conducted regarding the linguistic characteristics of retracted articles. For example, Markowitz and Hancock [11] evaluated the writing style of a single fraudulent author, social psychologist Diederik Stapel. Their finding showed that the author's writing style differed across his fraudulent and genuine articles. In a more recent study, Markowitz and Hancock [12] compared the writing style of 253 fraudulent articles (retracted for fraudulent data) partially matched with 253 non-fraudulent articles based on year, journal and keywords. These articles were indexed in PubMed and published in the time period of 1973–2013. They used an obfuscation index

Corresponding author:

Tahereh Dehdarirad, Department of Communication and Learning in Science, Chalmers University of Technology, Hörsalsvägen 2, Gothenburg 412 96, Sweden.

Email: tahereh.dehdarirad@chalmers.se

to do this. Their results showed that fraudulent articles tend to demonstrate a higher rate of linguistics obfuscation. The findings also showed a lower rate of positive emotion terms and a higher rate of causal terms in fraudulent articles.

This study also follows in the same vein and expands the latter line of research by applying exact matching on year, journal and topic and covariance balancing on several important factors, such as authors' gender, country and scientific impact. It additionally uses the weights obtained from matching and covariate balancing in different regression models, to provide a more accurate comparison between retracted and non-retracted articles in terms of language use. Language use in this article was measured from the point of view of use of positive terms and certainty language. (Un)Certainty is an essential components of science communication and presenting uncertainty in scientific works might influence reader's perception of scientific findings and their trust in science [13,14]. In biomedical research, the certainty or uncertainty of information communicated by authors in scientific articles through a series of linguistic markers plays a significant role in determining whether that information will be translated into practice or not [15]. Given the importance of certainty in science reporting and previous research that has showed language cues vary regarding use of positive terms in retracted and non-retracted articles, this study aimed to compare these two groups of articles in biochemistry, in terms of proportion of positive terms, certainty score and different certainty aspects.

To achieve the aim of study, the two research questions have been addressed below:

- Is there any difference in the proportion of positive terms used in the in abstracts, titles and findings presented in conclusion and discussion of retracted and non-retracted articles?
- Is there any difference between retracted and non-retracted articles in terms of uncertainty language (certain probability aspect, certain framing, certain suggestion) when findings are reported?

2. Data collection and processing

2.1. Data set

All journal articles published in time period of 2018–2020 in the subject area of biochemistry were downloaded from Scopus. These downloaded articles accounted for 331 retracted and 806,255 non-retracted articles. These 331 articles were retracted due to scientific misconduct, either in relation to issues with data or findings. After matching and covariate balancing analysis, 331 of 806,255 non-retracted articles were matched with 331 retracted articles. This accounted for a total of 662 articles, which was used for analysis in this article. Details regarding matching and covariate balancing can be found in the 'Data analysis' section. The reason for choosing biochemistry was previous research that found this field to be among the subject areas with a high number of retracted articles [9,16]. Choosing this time span allowed to consider a 3-year time window for possible retraction. Previous research works conducted in biomedical literature found a varied mean range of 26 [5] and 32.91 months [17] from publication to retraction for an article.

2.2. Outcome and covariates

In this section, we provide details regarding data collection and processing of the variables used in matching and covariate balancing as well as regression analysis.

2.2.1. Outcome variables. In model 1, the proportion of positive terms for each article was calculated by dividing the sum of positive terms appearing in the abstract, title and findings presented in conclusion and discussion by the length of the abstract and title (calculated as the number of words). We only included those sentences from discussion and conclusion sections where the authors were describing their own findings. As there were only 662 articles, we manually extracted these sentences. To avoid duplication, we also removed sentences from our analysis that were already mentioned in the abstracts.

Stop words were not considered in the calculation of the length of abstracts and titles. Python was used to do the data processing for natural language processing. We used a list of 25 positive terms that previous research had identified as positive in the titles and abstracts of biomedical articles [18,19]. For each positive term, a manual check was conducted to assure that positive words were not negated. For example, encouraging can appear as 'not encouraging'. If they were, we did not count them as positive terms and thus they were eliminated.

For models 2–5, the outcome variables were certainty score and certainty aspects (certain probability, certain framing and certain suggestion).

Certainty score and certainty aspects were calculated using the findings in the conclusion and discussion sections of the 662 articles. The Python certainty-estimator package was used to do this. Certainty-estimator is a package for

Table 1. Certainty aspect levels, their corresponding definition and an example sentence.

Aspect	Definition	Example sentence
Number	Refers to certainty towards specific quantities	'Our systematic search yielded 32 missense mutations' (certain)
Extent	Refers to certainty about the proportion/ratio of properties that make up an object/event or the extent of a change	'In this study level of adherence to insulin therapy was 59.9%' (certain)
Probability	Refers to certainty about the probability that something will occur, has occurred, or is associated with another factor	'This study showed that the downregulation of miR-143-5p contributed to the odontoblast differentiation of DPSCs' (certain)
Framing	Refers to the certainty about how scientists frame or interpret the scientific finding	'In summary, we found that miR-145-3p overexpression is able to inhibit the migration' (certain)
Condition	Refers to the situation where something depends on a specific condition, and the condition involves certainty or uncertainty	'Further studies are required to clarify the effect of other TUG1 targets in MM if any synergic relationship exists' (certain)
Suggestion	Refers to certainty or uncertainty about the implications or future actions for the public or science community	'Thus, further studies that allow us to understand the role of cyclin D1 on human OA chondrocytes are required to be discovered' (certain)

MM: Multiple myeloma; OA: Osteoarthritis.

estimating the certainty of scientific findings. The model in the package was trained over findings from diverse scientific domains, including biochemistry. The package provides two options to study certainty: sentence level and aspect level. The certainty score is between 1 and 6. The certainty aspects of scientific findings are categorised in: number, extent, probability, framing, condition and suggestion [20]. Table 1 shows these six different aspects, their definition and an example sentence for each aspect. The definitions are taken from Pei and Jurgens [20], whereas the examples are from the findings of the 662 studied articles.

Certainty score was calculated at sentence level. For each article, an average of certainty score was calculated based on all sentences reporting the findings in the conclusion section. In this article, we only compared retracted and non-retracted articles in terms of certain probability count, certain framing and certain suggestion aspects. These certainty aspects were the top 3 aspects that appeared in the findings of 662 studied articles (see 'Results', certainty-level aspects section).

2.2.2. Covariates. The covariates included in regression models affect both treatment and outcome and are called confounders. In this section, we have provided details regarding data collection and reasons regarding the inclusion of these covariates in the models.

To detect the gender of first and last authors, a combination of gender application programming interface (API) (<https://gender-api.com/>) and manual checking was carried out. First, gender API was used to conduct a search using the first name of authors. 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 scientific age of first and last authors was calculated using the geometric mean of citations. Both first and last authors were divided in quartiles based on their geometric mean of citations. In the models, they were entered as two categorical variables named first author impact and last author impact. The number or log-transformed number of publications and citations of an author has previously been defined as professional or scientific age of an author [21,22]. The reason for controlling for gender was that previous research found gender differences in the writing style of scientists and scientific discourse [23,24]. In addition, in terms of retraction, previous research found different patterns in terms of retraction between female and male scholars [25].

For topic, a categorical variable was created, which grouped articles into 20 topics. To do this, we used the classification data that were provided by Sjögarde [26]. This classification is based on 19 million PubMed publications from 1995 onwards and their citation relations. Labels in the classification have been created by extracting noun phrases from titles, author keywords, medical subject headings (MeSHs), journals and author addresses. Before performing covariate balancing and matching, there were 22 topics in both sets of retracted and non-retracted articles. However, after performing it, two topics were removed. Previous research has shown that different subjects can be differentiated by their writing styles [27]. Furthermore, some subject areas, such as chemistry [1] or cellular biology [9], have been found to have a higher rate of retraction. Regarding high-impact institute, a binary variable was created, which showed whether a top rank university existed in the affiliation list on a article (1) or not (0). In this study, by top rank we mean those universities that according to the Times Higher Education ranking are ranked as group 1 (1–200). The reason for inclusion of this variable

Table 2. Outcome, independent variables and covariates for matching and covariate balancing and regression analyses.

Variable type	Name	Measure
Outcome		
Model 1	Proportion of positive terms	
Model 2	Certainty score	
Model 3	Count of certain probability aspect	
Model 4	Count of certain framing aspect	
Model 5	Count of certain probability suggestion	
Treatment (independent)	Retraction	If article was retracted (1); otherwise (0)
Covariate	Journal ID (source ID)	Identifier for a journal in Scopus
	Publication year	Years 2018, 2019, 2020
	Topic	20 topics labelled from 0 to 20
	Gender of first author	Gender of first author on an article, male (0); Female (1)
	Gender of last author	Gender of last author on an article, male (0); Female (1)
	Author impact for first and last authors	Authors were divided in four quartiles based on their geometric mean of citation. The geo mean of citations was calculated for all articles published by first and last authors on an article based on data from Scopus database.
	Number of authors	Number of authors collaborating in an article.
	Native-English-speaking country	Whether any authors on an article was affiliated with an English-native-speaking country (1); or not (0)
	Open access	If articles were OA (1) or non-OA (0)
	High-impact institute	Whether any affiliation institutes on an article was a high-ranking institute (1); or not (0), this was based on the Times Higher Education ranking.

OA: Open Access.

in the model, was that previous research found universities with a higher ranking tend to have a lower rate of retraction [28]. Regarding country, a binary variable was created, which showed whether authors on a article were affiliated with a country where English is the official majority language (Australia, New Zealand, United Kingdom, Ireland, Canada and the United States) or had an affiliation outside these countries. Previous research has shown differences regarding the use of positive words between authors affiliated with these two groups of authors [18]. In addition, differences have been found between English and non-English languages, such as Bulgarian, in terms of certainty language (use of boosters and hedges) in academic texts [29].

Regarding the number of authors, according to previous research, having a large number of researchers may mitigate scientific errors and result in better reporting of studies, which may therefore avoid future retraction of articles [30]. In addition, some research works have found differences in readability and language use in single authored and co-authored articles [31].

Finally, we used Scopus to determine the open access status of articles. According to previous research, higher visibility of open access articles increases their readership and consequently the probability of detection of flaws in these publications [32].

Table 2 shows the covariates, outcome and independent variables in each regression model and in covariate balancing and matching analysis.

3. Data analysis

3.1. Matching and covariate balancing

To provide a more accurate comparison between retracted and non-retracted articles in terms of language use with the current data, we combined regression analysis with propensity score matching. This approach is referred to as ‘doubly robust estimator’ [33]. This helps to isolate causal factors by removing biases associated with differences between retracted and non-retracted articles. When regression and propensity score methods are used individually to estimate a

causal effect, they are unbiased only if the statistical model is correctly specified. The doubly robust estimator combines these two approaches, such that, only one of the two models needs to be correctly specified to obtain an unbiased effect estimator [33].

Using MatchIt R library, we did exact matching on three covariates of year, journal and topic. For journals, ISSN was used to do exact matching. Exact matching is a form of stratum matching that involves creating subclasses based on unique combinations of covariate values and assigning each unit into their corresponding subclass, so that, only units with identical covariate values are placed in the same subclass [34]. As an example, exact matching for a retracted article based on a journal means that a retracted article was matched to a non-retracted article, if they were both published in the same journal (using ISSN).

Exact matching on journal accounted for all level 2 variables related to a journal, such as its specialty or impact that might be correlated with the use of positive terms or certainty language. Bibliometric data have a multilevel structure. As articles are published in journals, journal is level 2 and article is level 1. Previous research has shown a positive association between the use of positive terms and being a high-impact factor clinical journal [19]. Doing exact matching on year and topic accounted for the possibility that positive presentation of research findings might vary over time or across different topics.

Covariate balancing for retracted and non-retracted was conducted on gender of first author, gender of last author, impact of first author, impact of last author, number of authors, English-native-speaking country, open access and high-impact institute. Using exact matching and covariate balancing, it was possible to gain the benefits of both exact matching and propensity score matching.

3.2. Regression models

To address the research questions of this study, five regression models were fitted. Model 1 compares retracted and non-retracted articles in terms of proportion of positive terms, whereas models 2–5 compare them in terms of certainty score and certainty aspects (certain probability, certain framing and certain suggestion).

3.2.1. Model 1: comparison of retracted and non-retracted articles in terms of proportion of positive terms used. As the dependent variable in this model is a proportion, a logistic regression with a binomial distribution was fitted. The logistic regression model is arguably the best model for proportion data when it is computed as ‘ ny out of n ’ with $Y = n_y/n$. The expected proportion is then modelled as binomial, where the explanatory variables contribute to its prediction through use of a logit link function [35]. Model quality was checked in terms of multicollinearity using variance inflation factor (VIF) test in *car* R package and binned residuals [36] using *performance* R package. See the supplementary material (Appendix 2, Model 1) for details.

3.2.2. Model 2: comparison of retracted and non-retracted articles in terms of certainty score. As the dependent variable in this model is a score, a linear regression was fitted. After checking the assumptions associated with linear regression models (linearity, normality of residuals, multicollinearity and heteroscedasticity), it was realised that the regression residuals suffered from heteroscedasticity ($p < 0.001$). According to previous research, continuous bounded outcome data that take values in a finite, typically display heteroscedasticity [37]. This is the case in our study as well, where the certainty scores take a value in the open interval of (2.68, 5.22). Beta regression has been proposed and judged suitable for modelling bounded outcome variables [38,39]. Beta distribution assumes values on the standard unit interval (0, 1), and it is flexible to model unimodal and bimodal data that are symmetric or skewed. If the variable takes on values in (a, b) (with $a < b$ known), then the response can be transformed to the (0, 1) interval by $(y - a)/(b - a)$, where b and a are the maximum and minimum possible scores [37,39], respectively. For details regarding model diagnostics, please see the supplementary material (Appendix 2, model 2).

3.2.3. Models 3 and 5: comparison of retracted and non-retracted articles in terms of count of certain probability aspect and certain suggestion aspect. As the dependent variables in both models are count variables, two Poisson regression models were fitted. The models fit were assessed using *DHARMA* package [40] in terms of dispersion, zero-inflation, outliers, collinearity and homogeneity of variances (see the supplementary material, Appendix 2, Models 3 and 5).

3.2.4. Model 4: comparison of retracted and non-retracted articles in terms certain framing aspect. As the dependent variable in this model is a count variable, a Poisson regression model was fitted. However, after checking the Poisson model using

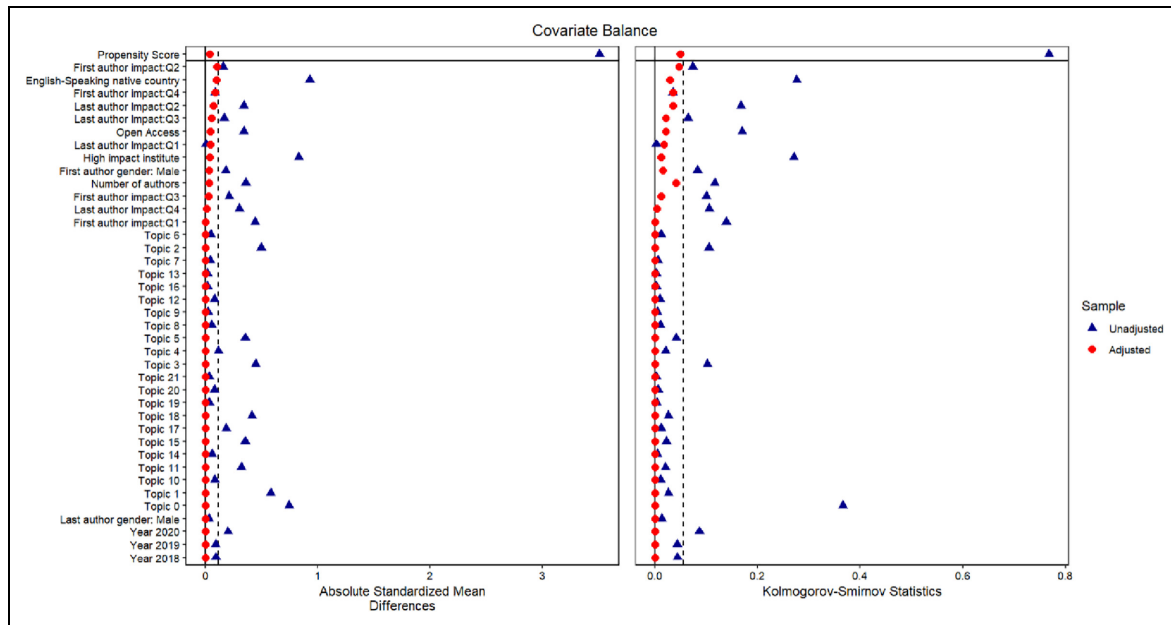


Figure 1. Covariate balance plot unadjusted versus adjusted mean difference and Kolmogorov–Smirnov statistic; horizontal lines mark critical values 0.1 and 0.05, respectively.

testDispersion() commands in R *DHARMA* package, it was realised that the model had under-dispersion. To deal with this, a quasi-Poisson model was fitted [41,42].

4. Results

4.1. Matching and covariate balancing

Figure 1 shows the absolute mean difference and Kolmogorov–Smirnov statistics for the adjusted and unadjusted case. As can be seen from the figure, the propensity scores matching weights substantially improve balance across all variables. These weights were used in the regression analyses.

4.2. Frequency of top 5 terms in retracted and non-retracted articles

Figure 2 shows the frequency of top 5 positive terms in retracted and non-retracted articles. As can be seen from the figure, except for term ‘interesting*’ and ‘support*’, which both had very similar frequency, non-retracted articles used terms important, novel and effective more frequently in comparison with retracted articles. However, the results of two sample proportion test showed that, only terms ‘important*’ and ‘novel’ were used more significantly in non-retracted articles ($p < 0.001$). Among the 25 positive terms studied, the only term that had a significantly higher frequency in retracted articles was ‘remarkabl*’ (remarkable, remarkably) with frequency of 70, in comparison with 50 for non-retracted articles ($p = 0.01$).

4.3. Regression results

In the regression tables in this section, the results were reported only for the variable retraction. The full result of the regression analyses for all covariates can be consulted in the supplementary material, Appendix 1 (Tables S1–S5).

4.3.1. Comparison of retracted and non-retracted articles in terms of proportion of positive terms. As can be seen from the table, retracted articles were 16% less probable to use positive terms in abstracts, titles, conclusion and findings sections in comparison with non-retracted articles (Table 3).

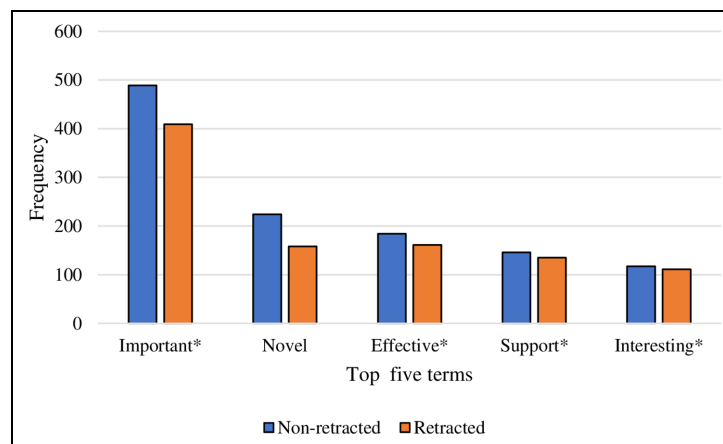


Figure 2. The frequency of top 5 positive terms in retracted and non-retracted articles.

Table 3. The result of logistic regression for comparison of retracted and non-retracted articles in terms of proportion of positive terms.

Predictors	Model 1: proportion of positive of terms		
	Estimates	Odds ratios	CI
Intercept	− 5.84	0.00***	0.00–0.00
Retracted (yes)	− 0.17	0.84***	0.78–0.91
N	662		
Pseudo R ²	0.40		

CI: confidence interval.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

4.3.2. Comparison of certainty language

Certainty score. Regarding certainty score, as can be seen from the table, the results showed that retracted articles had 15% lower odds in terms of use of certainty language, measured as certainty score, in presenting their scientific findings. As explained in the methodology section, these scientific findings were contained in the conclusion sections of the articles (Table 4).

4.3.2.1. Certainty aspect-levels: retracted vs non-retracted articles. Figure 3 compares retracted and non-retracted articles in terms of six certainty-level aspects. As can be seen from the figure, certain probability, certain framing and certain suggestion were the top 3 certain aspects appeared in the findings of 662 studied articles. Among these three certain aspects, except for certain probability that was used more frequently in the representation of findings of non-retracted articles, certain framing and certain suggestion were used more frequently with retracted articles. Regarding the uncertainty aspects, uncertain probability was used more frequently in the presentation of findings for retracted articles than non-retracted articles.

Table 5 shows the results of Poisson regression analysis for comparison of retracted and non-retracted articles in terms of count of certain probability aspects. As can be seen from the table, being a retracted article was associated with 11% decrease in the count of certain probability aspects used to present the research findings. In other words, retracted articles had 11% less likelihood to present their research findings using certain probability aspects.

Table 6 shows the results of quasi-Poisson regression analysis for comparison of retracted and non-retracted articles in terms of count of certain framing aspects. As can be seen from the table, there was no significant difference between retracted and non-retracted articles in term of use of certain framing aspect to present the research findings.

Regarding count of certain suggestion aspects, as can be seen from Table 7, the results showed no significant difference between retracted and non-retracted articles in terms of count of certain suggestion aspects that were used when presenting their scientific findings.

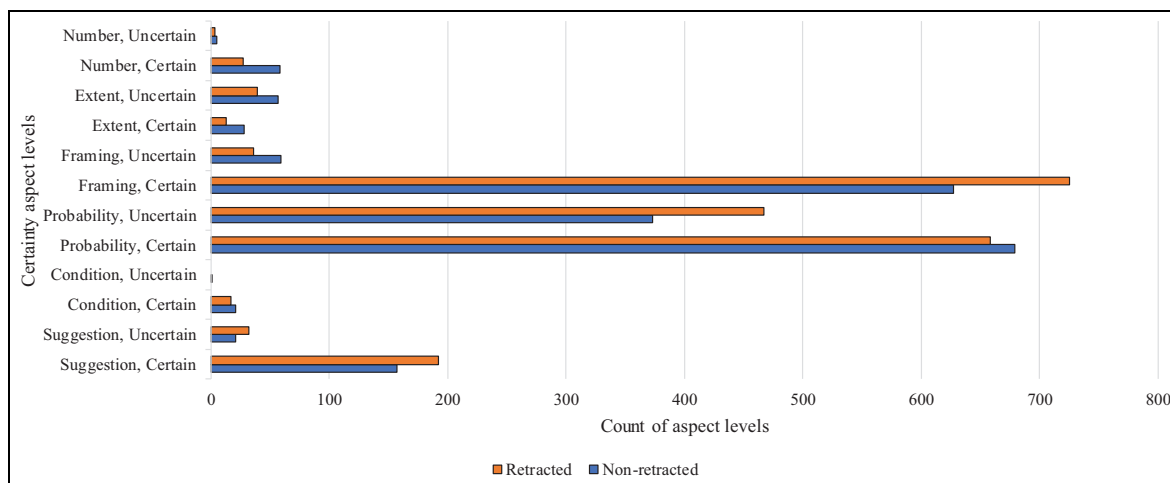


Figure 3. The comparison of retracted and non-retracted articles in terms of count of six certainty aspect levels.

Table 4. The result of beta regression for comparison of retracted and non-retracted articles in terms of certainty score.

Predictors	Model 2: certainty score		
	Estimates	Odds ratios	CI
Intercept	1.47	4.35***	1.88–10.09
Retracted (yes)	–0.16	0.85*	0.76–0.96
Observations	662		
R ²	0.20		

CI: confidence interval.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 5. The result of Poisson regression analysis for comparison of retracted and non-retracted articles in terms of count of certain probability aspect.

Predictors	Model 3: count of certain probability aspect		
	Estimates	Odds ratios	CI
Intercept	1.31	3.70***	1.86–7.38
Retracted (yes)	–0.12	0.89*	0.80–0.99
N	662		
R ² Nagelkerke	0.40		

CI: confidence interval.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

5. Conclusion and discussion

The aim of this study was to compare retracted and non-retracted articles in biochemistry, in terms of proportion of positive terms, certainty score and different certainty aspects (probability, framing and suggestion). Using a combination of regression analysis and covariate balancing and matching, this study attempted to remove biases associated with differences between retracted and non-retracted articles. This technique, which was previously mentioned as a ‘doubly robust’ estimator, has been recommended in bibliometrics studies as a technique to strengthen the robustness of the results [43]. Our analyses yielded several findings that are briefly presented and discussed below.

Table 6. The result of quasi-Poisson regression analysis for comparison of retracted and non-retracted articles in terms of count of certain framing aspect.

Predictors	Model 4: count of certain framing aspect		
	Estimates	Odds ratios	CI
Intercept	1.15	3.17***	1.78–5.43
Retracted (yes)	0.06	1.06	0.96–1.16
N	662		
R ² Nagelkerke	0.30		

CI: confidence interval.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.**Table 7.** The result of Poisson regression analysis for comparison of retracted and non-retracted articles in terms of count of certain suggestion aspect.

Predictors	Model 5: count of certain suggestion aspect		
	Estimates	Odds ratios	CI
Intercept	– 0.11	0.90	0.27–2.96
Retracted (yes)	0.15	1.16	0.94–1.45
N	662		
R ² Nagelkerke	0.30		

CI: confidence interval.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The findings with regards to the analysis of positive terms, showed that the term ‘remarkabl*’ was the only term among the 25 studied terms, which was used more frequently in the titles, abstracts and findings presented in ‘Conclusion and discussion’ sections of retracted articles. Among the top 5 terms, ‘important*’, ‘novel’ and ‘effective*’ were more frequently used in non-retracted articles. However, only ‘important*’ and ‘novel’ were used more significantly in non-retracted articles.

The results of regression analysis regarding the proportion of positive terms showed that retracted articles had lower odds of 16% to use positive terms in abstracts and titles in comparison with non-retracted articles. This finding is in line with Markowitz and Hancock [12] study, which also found a lower rate of positive emotion terms in retracted articles. The authors of retracted articles may use fewer positive terms not to draw attention to their research, which could lead to closer investigation of their work [44].

The findings regarding the analysis of certainty language, showed that retracted articles were 15% less probable to use certain language (measured by certainty score) when presenting their scientific findings. Uncertainty is a normal and necessary characteristic of scientific work as science involves producing knowledge about what was previously unknown [45]. However, in line with previous research [12], it might be possible that authors of retracted articles used an uncertain language to make their findings less comprehensive, thereby masking their deception. Regarding the three studied certainty aspects, the findings showed that except for certain probability, which was used more frequently in the presentation of findings by non-retracted articles, certain framing and certain suggestion were used more frequently with retracted articles. The results of further analysis by regression models showed that retracted articles had 11% less likelihood to present their research findings using certain probability aspect. The reason for this could be that authors of retracted articles might lack knowledge, data, methods or consciousness of aspects of a problem and, therefore, cannot be certain and precise in presenting and interpreting their findings. In their study, Funtowicz and Ravetz [46] referred to this type of scientific uncertainty as a ‘border with ignorance’, which cannot be sufficiently quantified and expressed with statistical measures. Our findings are in line with the work of Mehta and Guzmán [47], who in a slightly different context found that news outlets used probabilistic language, such as ‘may’, ‘likely’ and ‘possible’, in an uncertain way to add nuance to statistical discussions thereby making damaging claims more authentic.

Overall, the findings from our research suggest that linguistic characteristics can serve as a tool for distinguishing between retracted and non-retracted articles. They can also assist us in better understanding of how scientific misconduct

in scientific articles can affect communication and presentation of scientific findings. These findings also provide insight and guide policy for journal editors and authors within biochemistry discipline. In addition, better understanding of linguistic features of retracted and non-retracted could be helpful for classification of retracted articles (due to misconduct) and non-retracted articles using machine learning algorithms. Finally, the current research has some limitations. The findings obtained from this study are only limited to the field of biochemistry and certain time span and thus are not generalisable to other disciplines. However, it would be interesting to do a follow-up study in the future and compared those results with the ones obtained in this study. In addition, in this study we only studied certainty aspects of language and use of positive terms, other linguistic characteristics of retracted and non-retracted articles could also be studied in the future research.


Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Tahereh Dehdarirad  <https://orcid.org/0000-0003-2529-962X>

Marco Schirone  <https://orcid.org/0000-0002-4166-153X>

Supplemental material

Supplemental material for this article is available online.

References

- [1] Zhang M and Grieneisen ML. The impact of misconduct on the published medical and non-medical literature, and the news media. *Scientometrics* 2013; 96(2): 573–587.
- [2] Christopher MM. Comprehensive analysis of retracted journal articles in the field of veterinary medicine and animal health. *BMC Vet Res* 2022; 18(1): 73.
- [3] Hu G and Xu S. Agency and responsibility: a linguistic analysis of culpable acts in retraction notices. *Lingua* 2020; 247: 102954.
- [4] Bozzo A, Bali K, Evaniew N et al. Retractions in cancer research: a systematic survey. *Res Integr Peer Rev* 2017; 2(1): 5.
- [5] Cassão BD, Herbella FAM, Schlottmann F et al. Retracted articles in surgery journals. What are surgeons doing wrong? *Surgery* 2018; 163(6): 1201–1206.
- [6] Chauvin A, De Villelongue C, Pateron D et al. A systematic review of retracted publications in emergency medicine. *Eur J Emerg Med* 2019; 26(1): 19–23.
- [7] Al-Ghareeb A, Hillel S, McKenna L et al. Retraction of publications in nursing and midwifery research: a systematic review. *Int J Nurs Stud* 2018; 81: 8–13.
- [8] King EG, Oransky I, Sachs TE et al. Analysis of retracted articles in the surgical literature. *Am J Surg* 2018; 216(5): 851–855.
- [9] Serghiou S, Marton RM and Ioannidis JPA. Media and social media attention to retracted articles according to Altmetric. *PLoS ONE* 2021; 16(5): e0248625.
- [10] Stavale R, Ferreira GI, Galvão JAM et al. Research misconduct in health and life sciences research: a systematic review of retracted literature from Brazilian institutions. *PLoS ONE* 2019; 14(4): e0214272.
- [11] Markowitz DM and Hancock JT. Linguistic traces of a scientific fraud: the case of Diederik Stapel. *PLoS ONE* 2014; 9(8): e105937.
- [12] Markowitz DM and Hancock JT. Linguistic obfuscation in fraudulent science. *J Lang Soc Psychol* 2016; 35(4): 435–445.
- [13] Gustafson A and Rice RE. The effects of uncertainty frames in three science communication topics. *Sci Commun* 2019; 41(6): 679–706.
- [14] van der Bles AM, van der Linden S, Freeman ALJ et al. The effects of communicating uncertainty on public trust in facts and numbers. *Proc Nat Acad Sci* 2020; 117(14): 7672–7683.
- [15] Omero P, Valotto M, Bellana R et al. Writer's uncertainty identification in scientific biomedical articles: a tool for automatic if-clause tagging. *Lang Resour Eval* 2020; 54(4): 1161–1181.
- [16] Bhatt B. A multi-perspective analysis of retractions in life sciences. *Scientometrics* 2021; 126(5): 4039–4054.
- [17] Steen RG, Casadevall A and Fang FC. Why has the number of scientific retractions increased? *PLoS ONE* 2013; 8(7): e68397.

- [18] Vinkers CH, Tjldink JK and Otte WM. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *BMJ* 2015; 351: h6467.
- [19] Lerchenmueller MJ, Sorenson O and Jena AB. Gender differences in how scientists present the importance of their research: observational study. *BMJ* 2019; 367: l6573.
- [20] Pei J and Jurgens D. Measuring sentence-level and aspect-level (un)certainly in science communications. In: *The 2021 conference on empirical methods in natural language processing (EMNLP)*, 2021. Dominican Republic: Association for Computational Linguistics, <https://par.nsf.gov/servlets/purl/10336737>
- [21] Mishra S, Fegley BD, Diesner J et al. Self-citation is the hallmark of productive authors, of any gender. *PLoS ONE* 2018; 13: e0195773.
- [22] Andersen JP, Schneider JW, Jaggi 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.
- [23] Ma Y, Teng Y, Deng Z et al. Does writing style affect gender differences in the research performance of articles? An empirical study of BERT-based textual sentiment analysis. *Scientometrics* 2023; 128(4): 2105–2143.
- [24] Schmauss L and Kilian K. Hedging with modal auxiliary verbs in scientific discourse and women’s language. *Open Linguistics* 2023; 9(1): 20220229.
- [25] Decullier E and Maisonneuve H. Retraction according to gender: a descriptive study. *Account Res*. Epub ahead of print 13 October 2021. DOI: 10.1080/08989621.2021.1988576.
- [26] Sjögarde P. PubMed classification, 2022, <https://doi.org/10.6084/m9.figshare.c.5610971.v3>
- [27] Alluqmani A and Shamir L. Writing styles in different scientific disciplines: a data science approach. *Scientometrics* 2018; 115(2): 1071–1085.
- [28] Lievore C, Rubbo P, dos Santos CB et al. Research ethics: a profile of retractions from world class universities. *Scientometrics* 2021; 126(8): 6871–6889.
- [29] Dobakhti L. Expressing certainty in discussion sections of qualitative and quantitative research articles. *J Pan-Pac Assoc Appl Linguist* 2013; 17: 57–77.
- [30] Hartley J and Cabanac G. Are two authors better than one? Can writing in pairs affect the readability of academic blogs? *Scientometrics* 2016; 109(3): 2119–2122.
- [31] Li G, Kamel M, Jin Y et al. Exploring the characteristics, global distribution and reasons for retraction of published articles involving human research participants: a literature survey. *J Multidiscip Healthc* 2018; 11: 39–47.
- [32] Shah TA, Gul S, Bashir S et al. Influence of accessibility (open and toll-based) of scholarly publications on retractions. *Scientometrics* 2021; 126(6): 4589–4606.
- [33] Funk MJ, Westreich D, Wiesen C et al. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011; 173(7): 761–767.
- [34] Greifer N. Matching methods, 2022, <https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html>.
- [35] Chen K, Cheng Y, Berkout O et al. Analyzing proportion scores as outcomes for prevention trials: a statistical primer. *Prev Sci* 2017; 18(3): 312–321.
- [36] Gelman A and Hill J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press, 2006.
- [37] Ogundimu EO and Collins GS. Predictive performance of penalized beta regression model for continuous bounded outcomes. *J Appl Stat* 2018; 45(6): 1030–1040.
- [38] Kieschnick R and McCullough BD. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Stat Model* 2003; 3(3): 193–213.
- [39] Cribari-Neto F and Zeileis A. Beta regression in R. *J Stat Softw* 2010; 34: 1–24.
- [40] Hartig F. DHARMA: residual diagnostics for hierarchical (multi-level/mixed) regression models (R package version 03), 2020, <https://cran.r-project.org/web/packages/DHARMA/index.html>
- [41] Harris T, Yang Z and Hardin JW. Modeling underdispersed count data with generalized Poisson regression. *Stata J* 2012; 12(4): 736–747.
- [42] Musunuru A, Proffitt D, Ewing R et al. Poisson and negative binomial regression analysis. In: R Ewing and K Park (eds) *Advanced quantitative research methods for urban planners*. New York: Routledge, 2020, pp. 74–92.
- [43] Bittmann F, Tekles A and Bornmann L. Applied usage and performance of statistical matching in bibliometrics: the comparison of milestone and regular papers with multiple measurements of disruptiveness as an empirical example. *Quant Sci Stud* 2021; 2(4): 1246–1270.
- [44] Carey B. Stanford researchers uncover patterns in how scientists lie about their data. *Stanford News*, 16 November 2015.
- [45] Zehr SC. Scientists’ representations of uncertainty. In: Friedman SM, Dunwoody S and Rogers CL (eds) *Communicating uncertainty: media coverage of new and controversial science*. New York: Routledge, 1999, pp. 3–21.
- [46] Funtowicz SO and Ravetz JR. *Uncertainty and quality in science for policy*. Dordrecht: Kluwer Academic Publishers, 1990.
- [47] Mehta R and Guzmán LD. Fake or visual trickery? Understanding the quantitative visual rhetoric in the news. *J Media Lit Educ* 2018; 10(2): 104–122.