RESEARCH

CHRISTMAS 2015: THE PUBLICATION GAME

Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis

OPEN ACCESS

Christiaan H Vinkers *assistant professor*[1], Joeri K Tijdink *psychiatrist*[2], Willem M Otte *assistant professor*[3][4]

[1]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, 3584 CX Utrecht, Netherlands; [2]Department of Internal Medicine, VU University Medical Center, Amsterdam, Netherlands; [3]Department of Child Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht; [4]Biomedical MR Imaging and Spectroscopy, Center for Image Sciences, University Medical Center Utrecht, Utrecht

Abstract

Objective To investigate whether language used in science abstracts can skew towards the use of strikingly positive and negative words over time.

Design Retrospective analysis of all scientific abstracts in PubMed between 1974 and 2014.

Methods The yearly frequencies of positive, negative, and neutral words (25 preselected words in each category), plus 100 randomly selected words were normalised for the total number of abstracts. Subanalyses included pattern quantification of individual words, specificity for selected high impact journals, and comparison between author affiliations within or outside countries with English as the official majority language. Frequency patterns were compared with 4% of all books ever printed and digitised by use of Google Books Ngram Viewer.

Main outcome measures Frequencies of positive and negative words in abstracts compared with frequencies of words with a neutral and random connotation, expressed as relative change since 1980.

Results The absolute frequency of positive words increased from 2.0% (1974-80) to 17.5% (2014), a relative increase of 880% over four decades. All 25 individual positive words contributed to the increase, particularly the words "robust," "novel," "innovative," and "unprecedented," which increased in relative frequency up to 15 000%. Comparable but less pronounced results were obtained when restricting the analysis to selected journals with high impact factors. Authors affiliated to an institute in a non-English speaking country used significantly more positive words. Negative word frequencies increased from 1.3% (1974-80) to 3.2% (2014), a relative increase of 257%. Over the same time period, no apparent increase was found in neutral or random word use, or in the frequency of positive word use in published books.

Conclusions Our lexicographic analysis indicates that scientific abstracts are currently written with more positive and negative words, and provides an insight into the evolution of scientific writing. Apparently scientists look on the bright side of research results. But whether this perception fits reality should be questioned.

Introduction

Science has shown an impressive growth over past decades and more scientific papers are published now than ever before.[1] Between 1996 and 2011, over 15 million individuals authored around 25 million papers.[2] Owing to expanding research fields, it is increasingly difficult to get studies published in high impact journals.[3] This is important since publication quantity and associated impact factors have a considerable effect on a scientist's career perspective.[4] Consequently, in order to get published, scientific discoveries can sometimes be exaggerated or the potential implications overstated.[5][6] Indeed, overinterpretation, overstatement, and misreporting of scientific results have been frequently reported.[7-12] However, the prevalence of this problem in the scientific literature is unclear.

There is a well known universal tendency in people to use positive words,[13] and exaggeration of research related news has previously been linked to overstatements in academic press releases.[14] In the current study, we used a data driven approach to investigate trends in the use of positively and negatively valenced words in PubMed abstracts and titles over the past

Correspondence to: C H Vinkers c.h.vinkers@umcutrecht.nl

four decades. Subsequently, positive and negative word trends were compared with neutral and random words, as well as with patterns obtained from the corpus of digitised texts containing about 4% of all books ever printed, by use of Google Ngram Viewer. We hypothesised that the emergence of a culture aimed at productivity and novelty could have affected the use of positive and negative words in scientific reporting and discussion.

Methods

We quantified the yearly frequency of predefined positive, negative, and neutral words (25 words in each category) in titles and abstracts obtained from the PubMed database (box 1). Analyses were restricted to the period from 1 January 1974 to 31 December 2014, to ensure that all abstract texts were available. Words were selected after a consensus between the authors was reached through discussion, which included manual analysis of random abstracts and search of thesaurus listings. To validate the results from these pragmatic prespecified lists, we selected additional positive words from a recent article on superlatives in news coverage of cancer drugs.[15] To further exclude a bias in the choice of these words, we also searched for 50 nouns and 50 adjectives randomly selected from Ogden's 850 core words of Basic English.[16]

We divided the yearly number of abstracts containing one or more of the positive, negative, neutral, or random words in title or abstract text (based on the logical disjunction "or" operator) by the total number of yearly publications. The web appendix (web data 1) lists the search queries. We also summarised and tested differences between trends across the last 10 years in the search period using means and 95% confidence intervals and unpaired *t* tests. Patterns of individual words were plotted to determine whether developments were comparable across words. We calculated future predictions for the word "novel" with low order polynomial regressions. All analyses were carried out by use of R and plots were created with the R package ggplot2.

We needed to ensure that any trend in PubMed abstracts was specific for science rather than reflecting general trends in words used in society. Therefore, we also quantified the use of positive and negative words in books published between 1975 and 2009 using the Google Books Ngram Viewer, which charts the frequencies of words or short sentences in millions of books printed between 1800 and 2009.[17] We plotted average Google Books patterns and corresponding confidence intervals (calculated from bootstrap sampling of all individual word frequency patterns; 1000 samples/year) to evaluate differences with the patterns obtained from the PubMed queries.

In the light of the increasing number of journals and the rise of the open access movement, we selected (based on consensus) 20 journals with high impact factors that were likely to be of interest to biomedical readers (web appendix, web table S1). We deliberately did not include too many review based journals. We also examined two supplementary lists of journals: the top 20 journals listed in PubMed ranked by impact factor and the top 20 medical journals in PubMed, ranked by impact factor (both based on the Journal Citation Reports 2014; web table S1).

Finally, we investigated a possible cultural influence by comparing the use of positive and negative words in titles and abstracts between authors with an affiliation in a country where English is the official language (Australia, New Zealand, United Kingdom, and United States) and authors with affiliations outside these countries.

Results

Between 1974 and 1980, 1.7-2.3% of PubMed records containing one or more positive words in a title or abstract (mean 2.0% (standard deviation 0.21%)). This proportion increased to 17.5% in 2014, a relative increase of 880% (fig 1⇓, top left). Use of the same positive words in published books increased by 46% from 1975 to 2009 (fig 1⇓, top left).

Frequency patterns of all individual words in abstracts showed increased use during the search period, although with large variation (fig 2⇓). The words "robust," "novel," "innovative," and "unprecedented" increased in relative frequency from 2500% to 15 000% (fig 2⇓). Removal of these words still yielded a relative frequency increase of 540%. Moreover, word trends were similar after exclusion of low frequency words such as "inventive" and "astonishing." Analyses of additional positive words based on a recent article[15] ("breakthrough," "cure," "marvel," "miracle," "revolutionary," and "transformative") showed comparable and consistent patterns increases in frequency (web fig S1).

By extrapolating the upward trend of positive words over the past 40 years to the future, we predict that the word "novel" will appear in every record by the year 2123. Positive word use also increased in our selected list of journals with high impact factors, from 1.1% to 8.9% (relative increase 674%; fig 1⇓, top left). However, the increase in positive word use over the last 10 years of the search period was significantly lower in the group of our list of high impact factor journals than the frequency pattern of positive words across all journals (−159.8%, 95% confidence interval −92.9% to −226.7%, P<0.001). We found similar results in our analyses of the two supplementary lists of journals, ranked by impact factor (web fig S2).

Patterns in positive and negative words significantly differed between authors with an affiliation inside an English speaking country and those with affiliations outside an English speaking country. We saw reduced frequency rates in the last 10 years of the search period for authors affiliated with an institution in Australia, New Zealand, UK, or US compared to authors with an affiliation outside these countries (−31.4%, 95% confidence interval −50.6% to −12.2%, P=0.003; web fig S3).

For negative words, we saw a similar but less unequivocal increase in frequency. There was an absolute increase from 1.3% (standard deviation 0.07%) in 1974-80 to 3.2% in 2014, resulting in relative increases of up to 257%, and 199% if restricted to our selection of high impact journals (fig 1⇓, top right). Web figure 4 includes patterns of individual negative words. We saw no increase in the use of neutral words and only a modest increase in use for random words (fig 1⇓, bottom).

Discussion

Principal findings

Our analysis of scientific abstracts demonstrates that positive and—to a lesser extent—negative words are increasingly used over the past four decades. By contrast, this increase was absent for neutral and random words. The increase in positive words could not be attributed to general language tendencies as represented by the millions of printed books searched through in this study. Neither is the increase driven by one or two words, because all words showed increased frequency patterns. Even though the upward trend in positive word use was conserved in high impact journals, this trend was significantly less pronounced (fig 1⇓). This difference could be the result of a

---

Box 1: Words used in PubMed search queries and Google books search engine

Positive words

Amazing, assuring, astonishing, bright, creative, encouraging, enormous, excellent, favourable, groundbreaking, hopeful, innovative, inspiring, inventive, novel, phenomenal, prominent, promising, reassuring, remarkable, robust, spectacular, supportive, unique, unprecedented

Negative words

Detrimental, disappointing, disconcerting, discouraging, disheartening, disturbing, frustrating, futile, hopeless, impossible, inadequate, ineffective, insignificant, insufficient, irrelevant, mediocre, pessimistic, substandard, unacceptable, unpromising, unsatisfactory, unsatisfying, useless, weak, worrisome

Neutral words

Animal, blood, bone, brain, condition, design, disease, experiment, human, intervention, kidney, liver, man, men, muscle, patient, prospective, rodent, significant, skin, skull, treatment, vessel, woman, women

---

more thorough and critical editorial and peer review process in high impact journals.

Strengths and limitations of the study

The main strength of our lexicographic analysis was the inclusion of all PubMed abstracts over four decades, which would have prevented selection bias. Side by side comparisons with patterns of other word lists and general English texts provide robust reference data.

Our study also had limitations. Firstly, we limited the list of positive and negative words, and the choice of words was likely to have affected the specificity of the observed patterns. However, the general tendency was comparable across individual words, and sensitivity analyses with additional positive words yielded similar results. Secondly, we did not account for changes in the maximum abstract length of PubMed abstracts over the years. However, the upward trends are relatively linear over time, and abstract length would probably have resulted in an increase of neutral or random words as well.

Thirdly, we did not study the location of the words in the abstracts, or the context of their use. Contextual analysis of words could differ between the connotation of isolated words and the connotation conditional on the sentence. We also did not directly examine the association between word use and the current scientific culture—that is, the role of increased publication pressure and perceived relevance of publications for a scientific career. Finally, we cannot exclude the possibility that the scientific process has improved considerably over the past few decades and that the more frequent use of positive words is appropriate.

Implications of findings

Although it is possible that researchers have adopted an increasingly optimistic writing approach and are ever more enthusiastic about their results, another explanation is more likely: scientists may assume that results and their implications have to be exaggerated and overstated in order to get published. Our finding that scientific abstracts use more overt positive language is also probably related to the emergence of a positive outcome bias that currently dominates scientific literature.[18] There is much pressure on scientists in academia to publish as many papers as possible to further their careers. As a result, we may be afraid to break the bad news that many studies do not result in statistically significant or clinically meaningful effects.

Currently, most research findings could be false or exaggerated,[6 19] and research resources are often wasted.[20] Overestimation of research findings directly impairs the ability of science to find true effects and leads to an unnecessary focus on research marketability. This is supported by a recent finding that superlatives are commonly used in news coverage of both approved and non-approved cancer drugs.[21] The consequences of this exaggeration are worrisome since it makes research a survival of the fittest: the person who is best able to sell their results might be the most successful. It is time for a new academic culture that rewards quality over quantity and stimulates researchers to revere nuance and objectivity. Despite the steady increase of superlatives in science, this finding should not detract us from the fact we need bright, unique, innovative, creative, and excellent scientists.

1   Ridker PM, Rifai N. Expanding options for scientific publication: is more always better? *Circulation* 2013;127: 155-6. doi:10.1161/CIRCULATIONAHA.112.155952 23319809
2   Boyack KW, Klavans R, Sorensen AA, Ioannidis JP. A list of highly influential biomedical researchers, 1996-2011. *Eur J Clin Invest* 2013;43: 1339-65. doi:10.1111/eci.12171 24134636
3   Fraser AG, Dunstan FD. On the impossibility of being expert. *BMJ* 2010;341: c6815. doi:10.1136/bmj.c6815 21156739
4   Publish or perish. *Nature* 2015;521: 259. 25993920
5   Macleod MR, Michie S, Roberts I. Biomedical research: increasing value, reducing waste. *Lancet* 2014;383: 101-4. doi:10.1016/S0140-6736(13)62329-6 24411643
6   Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2: e124. doi:10.1371/journal.pmed.0020124 16060722
7   Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303: 2058-64. doi:10.1001/jama.2010.651 20501928
8   Ochodo EA, de Haan MC, Reitsma JB, et al . Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology* 2013;267: 581-8. doi:10.1148/radiol.12120527 23360738
9   Lockyer S, Hodgson R, Dumville JC, Cullum N. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. *Trials* 2013;14: 371. doi:10.1186/1745-6215-14-371 24195770
10   Patel SV, Chadi SA, Choi J, Colquhoun PH. The use of "spin" in laparoscopic lower GI surgical trials with nonsignificant results: an assessment of reporting and interpretation of the primary outcomes. *Dis Colon Rectum* 2013;56: 1388-94. doi:10.1097/01.dcr.0000436466.50341.c5 24201393
11   Boutron I, Altman DG, Hopewell S, et al . Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *J Clin Oncol* 2014;32: 4120-6. doi:10.1200/JCO.2014.56.7503 25403215

What is already known on this topic

Our perception of the world is reflected in how we use language

Science has shown a substantial growth over past decades, and in order to be published, scientific discoveries are sometimes exaggerated or potential implications overstated

What this study adds

Analysis of PubMed abstracts shows that positive words have been increasingly used between 1974 and 2014

Use of more overt positive language is probably related to the emergence of a positive outcome bias that currently dominates scientific literature

12 Lazarus C, Haneef R, Ravaud P, Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol* 2015;15: 85. doi:10.1186/s12874-015-0079-x 26462565

13 Dodds PS, Clark EM, Desu S. Human language reveals a universal positivity bias. *Proc Natl Acad Sci U S A* 2015;112: 2389-94. doi:10.1073/pnas.1411678112 25675475

14 Sumner P, Vivian-Griffiths S, Boivin J. The association between exaggeration in health related science news and academic press releases: retrospective observational study [correction in *BMJ* 2014;349:g7666]. *BMJ* 2014;349: g7015. doi:10.1136/bmj.g7015 25498121

15 McCarthy M. Superlatives are commonly used in news coverage of cancer drugs, study finds. *BMJ* 2015;351: h5803. doi:10.1136/bmj.h5803 26525225

16 Ogden's basic English. Homepage. 2012. http://ogden.basic-english.org/basiceng.html.

17 Michel JB, Shen YK, Aiden APGoogle Books Team. Quantitative analysis of culture using millions of digitized books. *Science* 2011;331: 176-82. doi:10.1126/science.1199644 21163965

18 Dwan K, Gamble C, Williamson PR, Kirkham JJReporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One* 2013;8: e66844. doi:10.1371/journal.pone.0066844 23861749

19 Ioannidis JP. How to make more published research true. *PLoS Med* 2014;11: e1001747. doi:10.1371/journal.pmed.1001747 25334033

20 Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009;374: 86-9. doi:10.1016/S0140-6736(09)60329-9 19525005

21 Abola MV, Prasad V. The use of superlatives in cancer research. *JAMA Oncol* 2015, doi: 10.1001/jamaoncol.2015.3931. 26512913
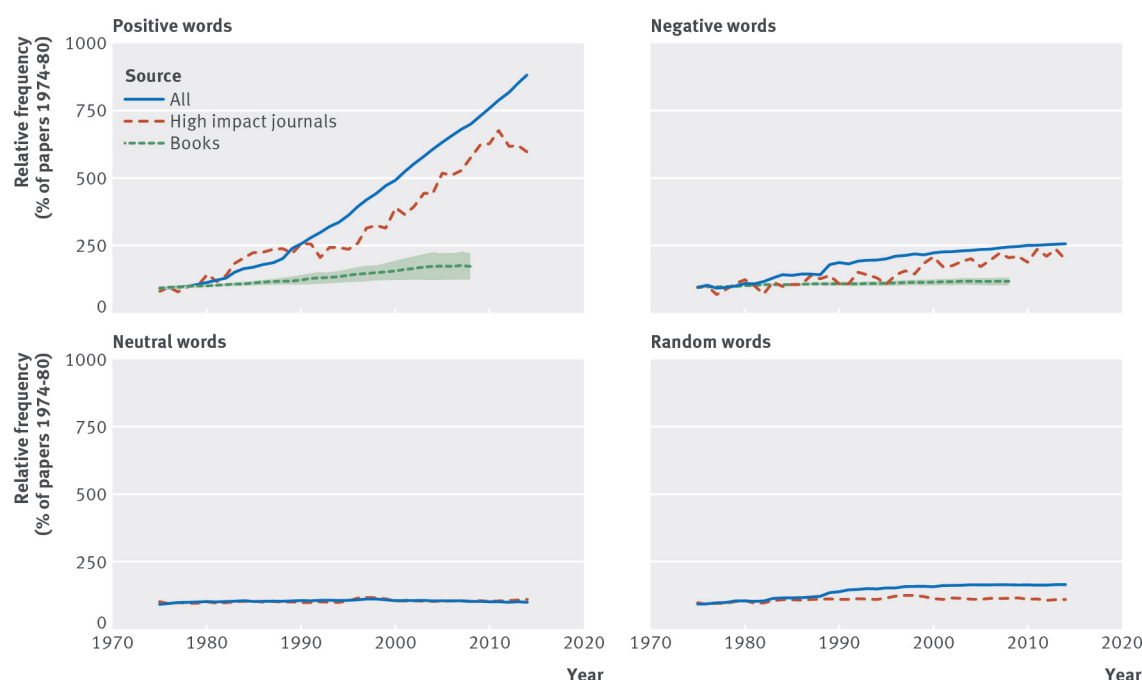
Figures

Fig 1 Relative frequency patterns of positive, negative, neutral, and random words in PubMed abstracts and titles over time. Mean relative frequency patterns of the same positive and negative words in general books is plotted with 95% confidence intervals (shaded grey). All=words in all abstracts; Books=words in digitalised Google Books corpus; High impact journals=words in our selected list of high impact factor journals
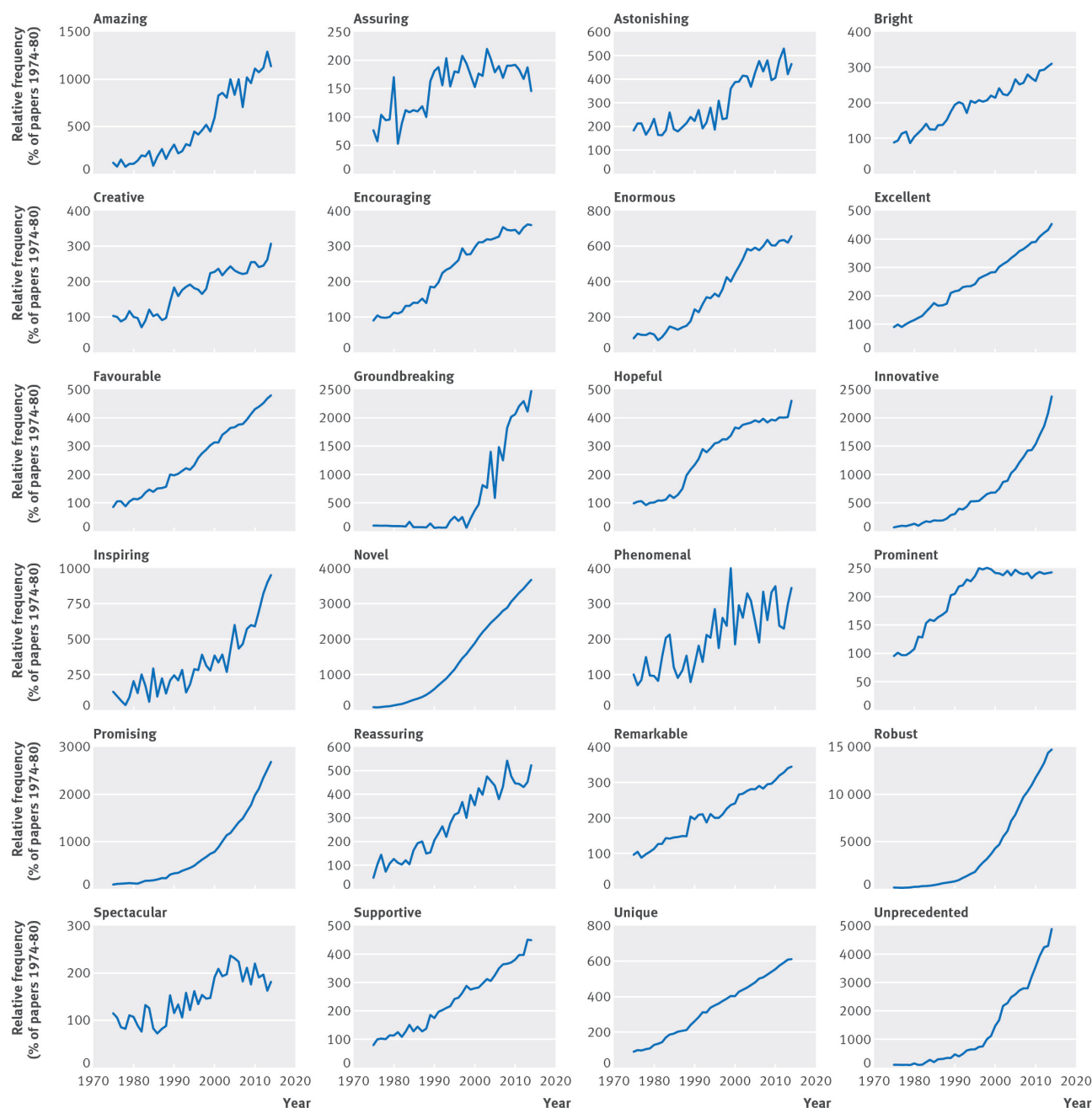
Fig 2 Relative frequencies of 24 individual positive words as used in PubMed between 1975 and 2014. The word "inventive" was not plotted owing to low search volumes