European Annals of Otorhinolaryngology, Head and Neck diseases 136 (2019) 469–473

Available online at ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com/en

Original article

Use of *P*-values and the terms "significant", "non-significant" and "suggestive" in Abstracts in the *European Annals of Otorhinolaryngology, Head & Neck Diseases*

O. Laccourreye[a,b,*], Q. Lisan[a,b], P. Bonfils[a,b], R. Garrel[a,c], R. Jankowski[a,d], A. Karkas[a,e], N. Leboulanger[a,b], M. Makeieff[a,f], C. Righini[a,g], C. Vincent[a,h], C. Martin[a]

[a] *Comité de rédaction, annales d'otorhinolaryngologie et de pathologie cervico-faciale, Elsevier ed., 92130 Issy-les-Moulineaux, France*
[b] *Université Paris Descartes Sorbonne Paris Cité, 75006 Paris, France*
[c] *Université de Montpellier, 34090 Montpellier, France*
[d] *Université de Lorraine, 54505 Nancy, France*
[e] *Université de Saint-Étienne, 42270 Saint-Étienne, France*
[f] *Université de Reims, 51100 Reims, France*
[g] *Université de Grenoble, 38700 Grenoble, France*
[h] *Université de Lille, 59120 Lille, France*

ARTICLE INFO

Keywords:
Medical writing
Scientific report
Statistics
P value
Significant
Suggestive
Spin.

ABSTRACT

Objectives: To evaluate the use of *P*-values and the terms "significant", "non-significant" and "suggestive" in Abstracts in the *European Annals of Otorhinolaryngology, Head & Neck Diseases*.
Materials and methods: Consecutive articles accepted for publication during the period January 2016 – February 2019 were systematically reviewed. Main goal: descriptive analysis of the citation of *P*-values and use of the terms "significant", "non-significant" and "suggestive" in Abstracts. Secondary goal: analytic study of: (i) correlations between citation of a *P*-value and the main characteristics of authors and topics; and (ii) misuse of the terms "significant", "non-significant" and "suggestive" with respect to cited *P*-values, and correlations with author and topic characteristics.
Results: In all, 91 articles were included. *P*-values and the terms "significant", "non-significant" and "suggestive" were cited in 35.1%, 41.7%, 10.9% and 0% of Abstracts, respectively. Citing a *P*-value did not significantly correlate with author or topic characteristics. There were discrepancies between the terms "non-significant", "significant" and "suggestive" and *P*-values given in the body of the article in 57.1% of Abstracts, with 30.7% overestimation and 25.2% underestimation of results, without significant correlation with author or topic characteristics.
Conclusion: Authors, editors and reviewers must pay particular attention to the spin resulting from inappropriate use of the terms "significant", "non-significant" and "suggestive" in Abstracts of articles submitted to the *European Annals of Otorhinolaryngology, Head & Neck Diseases*, to improve the rigor, quality and value of the scientific message delivered to the reader.

1. Introduction

The *European Annals of Otorhinolaryngology, Head & Neck Diseases* is the official journal of the French Society of Otorhinolaryngology (SFORL) and International French-Speaking Society of Otorhinolaryngology (SIFORL). The journal adheres to the guidelines of the International Committee of Medical Journal Editors (http://www.icm-je.org/), and has recently revised its Instructions to Authors [1].

The aim of the present study, in the light of the growing importance of statistics in biomedical articles and the increasing use of the terms "significant" and "suggestive" to characterize results in biomedical research [2–5], was to analyze the use of *P*-values and the terms "significant', "non-significant" and "suggestive" in the Abstracts of articles submitted to and accepted for publication in the *European Annals of Otorhinolaryngology, Head & Neck Diseases* for the period 2016–2019. The findings are discussed in the light of the literature on the place of Abstracts and the role of statistics in medical articles.

* Corresponding author.
E-mail address: ollivier.laccourreye@aphp.fr (O. Laccourreye).

https://doi.org/10.1016/j.anorl.2019.10.008
1879-7296/

Table 1
Typology (author and topic characteristics) of the 91 articles.

Characteristics of authors and topics	Number (%)
First author country	
France	73 (80.2)
China, Spain, Italy, Israel, Senegal, Turkey	2 (2.1)
Benin, South Korea, USA, Morocco, Czech Republic, Sweden, Tunisia	1 (1)
First author gender (female/male)	32 (35.1)/59 (64.9)
First author affiliation	
University hospital	82 (90.1)
Non-university hospital	4 (4.3)
Cancer center	3 (3.2)
Private practice	2 (2.1)
Number of authors per article (Fig. 1)	2 to 11 – median: 6 –
Article topic	
Otology	27 (29.6)
Rhinology	21 (23)
Laryngology	15 (16.4)
HNS/MF&PS	27 (29.6)/1 (1)
Article specificities	
Pediatrics	14 (15.3)
Oncology	27 (29.6)
Clinical/Fundamental	87 (96.7)/4 (4.3)
Prospective	7 (7.6)
Multicenter	9 (9.6)
Statistics unit	5 (5.4)

HNS: head and neck surgery; MF&PS: maxillofacial and plastic surgery.

2. Material and methods

Original articles submitted to and accepted by the Editorial Board for publication in regular issues of the *European Annals of Otorhinolaryngology, Head & Neck Diseases* for the period January 2016 through February 2019 were analyzed. Exclusion criteria comprised articles submitted before the study period, not accepted during the study period, accepted for publication in another section of the journal (literature review, review article, case report, technical note, editorial, tropical diseases, letter to the editor, or "What is your diagnosis?"), or submitted to and accepted for special issues published during the study period.

After double reading of the Abstracts and Results sections of articles included in the EEL database of articles submitted to and accepted for publication in the *European Annals of Otorhinolaryngology, Head & Neck Diseases* during the study period, the study comprised two parts. The first was descriptive, measuring the number of *P*-values cited in the Abstract and use of the terms "significant" ($P < 0.005$), "non-significant" ($P > 0.05$) and "suggestive" ($0.005 < P < 0.05$), in line with changes regarding significance thresholds recommended to improve the use of statistics in biomedicine [4,5]. The second part was analytic, assessing correlations between *P*-value citation in Abstracts and author and topic characteristics as detailed in Table 1. It then assessed whether the use of the terms "significant", "non-significant" and "suggestive" matched the *P*-values cited in the Abstract (or else in the Results section). Finally, it assessed correlations between misuse of the terms "significant", "non-significant" and "suggestive" and author and topic characteristics. Data were collected in a standardized computer file and analyzed on StatView software (StatView Inc. USA), using Fisher t test for qualitative variables and Mann-Whitney U test for quantitative variables. The significance threshold was set at $P < 0.005$ [2–5]. *P*-values > 0.05 and 0.05–0.005 were respectively designated "non-significant" and "suggestive" [4,5].

3. Results

Ninety-one articles were included. Table 1 details number of authors (see also Fig. 1), gender, country of origin, first-author health-care structure, topic and type of study.

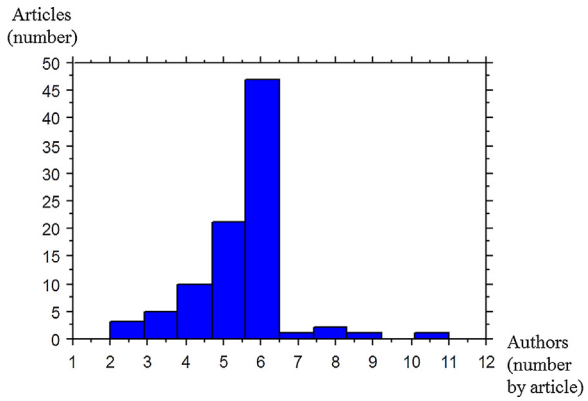

Fig. 1. Number of authors per article.

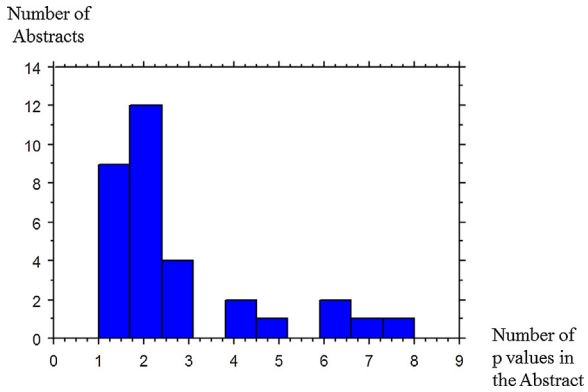

Fig. 2. Number of p-values in the 32 Abstracts with at least 1 *P*-value cited in the Abstract.

3.1. Descriptive study

Abstracts cited *P*-values and the terms "significant", "non-significant" and "suggestive" in respectively 35.1% (32/91), 41.7% (38/91), 10.9% (10/91) and 0% (0/91) of cases. The number of *P*-values cited in Abstracts citing at least 1 (Group A) ranged from 1 to 8 (median, 2; Fig. 2), with *P*-values ranging between < 0.0001 and 1.

The 59 articles with no *P*-values cited in the Abstract were of 2 types: those with *P*-values (median number, 5; range, 1–54), ranging from < 0.0001 to 1, cited in the Results section (Group B), and those with none (Group C). Groups B and C accounted for respectively 30.7% (28/91) and 34.1% (31/91) of articles.

3.2. Analytic study

On univariate analysis (Table 2), citing a *P*-value in the Abstract did not significantly correlate with any author or topic characteristics.

Misuse of the terms "significant", "non-significant" and "suggestive" in the light of the *P*-values cited in the Abstract (or else Results section) was found in 57.1% (52/91) of articles (Table 3): 75% (24/32) of articles in Group A, 92.8% (26/28) in Group B and 6.4% (2/31) in Group C, thus over- or under-estimating findings in respectively 30.7% and 25.2% of Abstracts.

On univariate analysis (Table 4), misuse of the terms "significant", "non-significant" and "suggestive" did not significantly correlate with any author or topic characteristics.

Table 2
Correlation between citation (Group A) or non-citation (Group B) of at least 1 *P*-value in the Abstract and author and topic characteristics.

Characteristics of authors and topics	Group A (at least 1 *P*-value in Abstract)	Group B (no *P*-value in Abstract)	*P*
First author country			
France (yes/no)	24/8	49/10	0.41
First author gender (f/m)	12/20	20/39	0.81
Academic first author (yes/no)	30/2	52/7	0.48
Number of authors per article			
Median (range)	6 (2–11)	6 (3/11)	0.59
Article topic			
Otology (yes/no)	8/24	19/40	0.63
Rhinology (yes/no)	6/26	14/45	0.79
Laryngology (yes/no)	9/23	6/53	0.038
HNS/MF&PS (yes/no)	9/23	20/39	0.64
Article specificities			
Pediatrics (yes/no)	4/28	10/49	0.76
Oncology (yes/no)	12/20	15/44	0.24
Clinical/fundamental	32/0	55/4	0.29
Prospective (yes/no)	3/29	4/55	0.69
Multicenter (yes/no)	3/29	6/53	0.99
Statistics unit (yes/no)	0/32	5/44	0.15

f: female; m: male; HNS: head and neck surgery; MF&PS: maxillofacial and plastic surgery.

Table 3
Typology of misuse of "non-significant", "significant" and "suggestive" in the Abstract; several misuses in a given Abstract are possible. (*n*: number).

	Misuses in Abstract	% (*n*)
Overestimation		30.7 (28/91)
	"significant" used but not "suggestive"	21.9 (20/91)
	"significant" used but not "non-significant"	6.5 (6/91)
	"significant" used instead of "suggestive"	8.7 (8/91)
	"significant" misused	2.2 (2/91)
	"non-significant" misused	3.2 (3/91)
Underestimation		25.2 (23/91)
	No term, whereas "suggestive" would be appropriate	20.8 (19/91)
	No term, whereas "significant" would be appropriate	12 (11/91)
	No term, whereas "non-significant" would be appropriate	6.5 (6/91)

Table 4
Correlation between misuse of "non-significant", "significant" and "suggestive" in the Abstract and author and topic characteristics.

Characteristics of authors and topics	Correct use	Misuse	*P*
First author country			
France (yes/no)	45/7	28/11	0.11
First author gender (f/m)	19/33	13/26	0.82
Academic first author (yes/no)	48/4	34/5	0.48
Number of authors per article			
Median (range)	6 (3–11)	6 (2–9)	0.26
Article topic			
Otology (yes/no)	16/36	11/28	0.82
Rhinology (yes/no)	13/39	7/32	0.45
Laryngology (yes/no)	10/42	5/34	0.57
HNS/MF&PS (yes/no)	13/39	16/23	0.11
Article specificities			
Pediatrics (yes/no)	6/46	8/31	0.25
Oncology (yes/no)	17/35	10/29	0.49
Clinical/fundamental	51/1	36/3	0.3
Prospective (yes/no)	6/46	1/38	0.23
Multicenter (yes/no)	5/47	4/35	0.99
Statistics unit (yes/no)	3/49	2/37	0.99

f: female; m: male; HNS: head and neck surgery; MF&PS: maxillofacial and plastic surgery.

4. Discussion

"Scientific", also known as "original", articles are the motor of biomedical research. Since the 1970s, they have used a 5-section structure to promote the concision, rigor and reproducibility of the scientific message, in a campaign led by the International Committee of Medical Journal Editors (http://www.icm-je.org/).

Among these five sections, the Abstract has a special role to play. With its limited word-count, it is the display window of the article, and authors, editors and readers all have a special interest for it. Based on their reading of the Abstract, the editors assess the originality of the study, its structure and compliance with the instructions to authors, and decide on its suitability for submission to peer review. The Abstract is moreover in open access on-line, available for ENT specialists to study in their literature searches to prepare research projects and to select articles to be examined in full text for possible inclusion in the reference list of their publications. And finally, the "open access tsunami" [6] means that Abstracts are the essential vehicle of medical information for the general public. Drawing up the Abstract is thus a key step in the rapid globalized publication of the message emerging from scientific research. And shortcomings in composing the Abstract are, in our opinion, one of the main reasons why 21% of the 39 million biomedical articles to be found on the Web of Science for the period 1990–2015 have been cited only once or not at all [7].

In everyday language, the term "abstract" suggests a selection, in the sense of a short collection of data drawn or "abstracted" from various sources, and the Abstract of a scientific article is a short structured summary of the article as a whole, highlighting the most relevant findings and key data, underlining their scientific interest and novelty and their implications for research and for patients. These basic features of Abstract composition mean that, little by little, *P*-values and the terms "significant' and "non-significant" have worked their way into the text, providing the study and the ideas it is promoting with the statistical back-up that has become a prerequisite in a world in which quantitative assessment occupies a key position in all realms of science.

Thus, in 2016, Chavalarias et al. [2], in an analysis of almost 13 million biomedical Abstracts available in the MEDLINE database, found that the percentage of articles citing *P*-values in their Abstract had risen from 7.3% in 1990 to 15.6% in 2014. In 2014, *P*-values were cited in the Abstracts of 38.9% of clinical trials and 54.8% of randomized controlled trials [2]. Likewise, in 2006, in an analysis of more than 2,500 articles, 75% of which reported clinical research published in 1993, 1998 and 2003 in 4 English-language ENT journals considered "major' by the authors, Waserman et al. [8] found that the rate of *P*-value citation increased from 26% to 45%. With a rate of 35.1% in Abstracts published between 2016 and 2019 in the *European Annals of Otorhinolaryngology, Head & Neck Diseases*,

regardless of author or topic characteristics (Table 2), the journal of the French Society of ENT seems well in line with current trends.

The aim of any research is to assess some hypothesis or other (e.g., that postoperative radiation therapy affects the rate of cure). To this end, two statistical hypotheses are formulated: one which statisticians call the "null hypothesis" (e.g., the cure rate after postoperative radiation therapy is identical to that without radiation therapy – or arithmetically, cure rate after RT minus cure rate without RT = 0, whence the term "null"), and one called "alternative" (e.g., the difference in cure rate between the two conditions is not zero). If the difference between the two study groups turns out not to be zero, the null hypothesis is rejected and the alternative hypothesis is accepted. If the cure rate in the group with postoperative radiation therapy is higher than in the group without, then statistical tests are needed to estimate the probability (*P*-value) of reaching such a result by pure chance, so as to know whether the findings are significant or not (http://www.cons-dev.org/elearning/stat/St2a.html).

Ronald Aylmer Fisher [9] was a British researcher, mathematician and geneticist, who died in 1962 and was considered by his peers [10] to be "Darwin's greatest successor", who almost alone founded modern statistics. It was he who suggested adopting a 5% *P*-value threshold in biomedicine for a result to be considered "significant" and not attributable to mere chance–or, in ordinary language, that there is less than 5% risk of being wrong in claiming that there is a difference between the two groups.

Almost a century after Fisher's work was published, a replication crisis in many biomedical studies using the term "significant" in case of *P*-values < 0.05 and the exposure of some very bad practices in conflict with all medical ethical standards have given rise to a debate which for several years has been stirring the statistical community, some members of which do not hesitate to talk about a perversion of statistics on the part of clinicians [3,4,11–13].

The question currently arising is this: should the *P*-value threshold be 0.005 instead of 0.05, with only lower values being considered "significant"? Lowering the threshold would reduce the number of results counting as "significant' and rule out some false positives. Results with *P*-values between 0.05 and 0.005 may be deemed "suggestive", justifying going ahead with the research while tightening it up, preferably with a multicenter randomized design and larger sample size [4,13,14]. With a 57.1% rate of misuse of the term "significant" in Abstracts submitted and published between 2016 and 2019 in the *European Annals of Otorhinolaryngology, Head & Neck Diseases* (Table 4), the present study argues for just such a change – especially as the term "suggestive" was not used at all and the term "non-significant" featured in only 10.9% of Abstracts, in line with the findings of Chavalarias et al. [2], who found that, in almost 13 million biomedical Abstracts, whenever *P*-values were actually cited they were almost always said to be "significant". In the present study, the terms "significant' and "non-significant" (Table 3) were misused in 34.1% of Abstracts; moreover, in a large number of cases the authors did not refer to *P*-values at all, despite the fact that they were cited in the Results section and could be considered significant at the 0.005 level or non-significant in respectively 12% and 9.8% of cases. This misuse of terms in the Abstract thus underplayed the impact of the scientific message in 25.2% of cases, or a quarter of articles, and overplayed it in 30.7%, or about one-third (Table 3). The fact that misuse of the terms "significant", "non-significant" and "suggestive" in the Abstract was unrelated to author or topic characteristics (Table 4) only goes to amplify the importance of the statistical problem revealed in the present study, suggesting, in our opinion, a failure of initial and continuous training and of information regarding the correct use of *P*-values. This is sufficiently important for the editors of the *European Annals of Otorhinolaryngology, Head & Neck Diseases* to have suggested setting up a post of statistics editor [1].

The authors, reviewers and editors of the *European Annals of Otorhinolaryngology, Head & Neck Diseases* also need to be aware that knowing the right threshold for *P*-values and how to use the terms "significant" and "suggestive" in an Abstract is just the public face of good statistical practice. If statistical analysis is to have any real scientific value, it is essential that the tests that are used (parametric, non-parametric, actuarial, correlation, etc.) are suited to the null hypothesis (http://www.cons-dev.org/elearning/stat/St2a.html) and to the study population, and that interpretation takes proper account of confidence intervals and effect size, using Bayesian methods as appropriate [15–17]. In 2017, in an analysis of 107 mainly retrospective studies consecutively published between 2012 and 2015 in the American journal *JAMA Otolaryngology-Head & Neck Surgery*, Karadaghy et al. [18] found that only 55% specified the impact of effect size when the result was considered worthy of interest. Also, in 2017, Consonni and Bertazzi [19] stressed the danger of lack of statistical rigor in an area that all ENT practitioners are bound to encounter at some point in their career: the medico-legal field.

5. Conclusion

To enhance the rigor, quality and value of the scientific information provided to readers of the *European Annals of Otorhinolaryngology, Head & Neck Diseases*, authors, editors and reviewers should pay greater attention to the use of the terms "significant" and "non-significant" in the Abstracts of submitted articles, in order to avoid overestimation of statistical results and spin [20], which are all too frequent in medicine.

Disclosure of interest

The authors declare that they have no competing interest.

Acknowledgments

The authors thank the association *Progrès 2000* for technical support.

References

[1] Laccourreye O, Lisan Q, Martin C. Statistics "officially" integrate the European Annals of Otorhinolaryngology Head & Neck Diseases. Eur Ann Otorhinolaryngol Head Neck Dis 2019, http://dx.doi.org/10.1016/j.anorl.2019.02.002 [S1879-7296(19)30016-X].
[2] Chavarralias D, Wallach JD, Li AH, et al. Evolution of reporting p values in the biomedical literature, 1990–2015. JAMA 2016;315:1141–8.
[3] Begley CG, Ioannidis JP. Reproducibility in science. Improving the standard for basic and preclinical research. Circ Res 2015;116:116–26.
[4] Ioannidis JPA. The proposal to lower p value thresholds to.005. JAMA 2018;319:1429–30.
[5] Laccourreye O, Maisonneuve H. Toward a new threshold for the P value? Eur Ann Otorhinolaryngol Head Neck Dis 2018;135:299.
[6] Maisonneuve H, Laccourreye O. The open access tsunami hits France. Eur Ann Otorhinolaryngol Head Neck Dis 2019, http://dx.doi.org/10.1016/j.anorl.2019.02.003 [S1879-7296(19)30017-1].
[7] Van Noorden R. The science that's never been cited. Nature 2017;552:162–4.
[8] Wasserman JM, Wynn R, Bash TS, Rosenfeld RM. Levels of evidence in otolaryngology journals. Otolaryngol Head Neck Surg 2006;134:717–23.
[9] Fisher RA. Statistical methods for research workers. Edinburgh: Oliver & Boyd ed.; 1925.
[10] Dawkins R. River out of Eden. A Darwinian Wiew of Life. New York: Basic Books ed.; 1995.
[11] Laccourreye O, Maisonneuve H. Suggestive" and "Significant": you can't always get what you want. . .. Eur Ann Otorhinolaryngol Head Neck Dis 2019, http://dx.doi.org/10.1016/j.anorl.2019.02.006 [S1879-7296(19)30031-6].
[12] Krishna A, Peter SM. Questionable research practices in student final theses - Prevalence, attitudes, and the role of the supervisor's perceived attitudes. PLoS One 2018;13(8):e0203470, http://dx.doi.org/10.1371/journal.pone.0203470 [eCollection 2018].
[13] Miller J, Ulrich R. The quest for an optimal alpha. PLOS One 2019;14(1):e0208631, http://dx.doi.org/10.1371/journal.pone.0208631.

[14] Lytsy P. P in the right place: revisiting the evidential value of p-values. J Evid Based Med 2018;11:288–91.
[15] Vila PM, Townsend ME, Bhatt NK, Kao WK, Sinha P, Neely JG. The p value problem in otolaryngology: shifting to effect sizes and confidence intervals. Otolaryngol Head Neck Surg 2017;156:978–80.
[16] Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: what do p values and confidence intervals really represent? Anesth Analg 2018;126:1068–72.
[17] Buchinsky FJ, Chadha NK. To p or not to p: backing Bayesian statistics. Otolaryngol Head Neck Surg 2017;157:915–8.
[18] Karadaghy OA, Hong H, Scott-Wittenborn N, et al. Reporting of effect size and confidence intervals in JAMA Otolaryngology-Head & Neck Surgery. JAMA Otolaryngol Head Neck Surg 2017;143:1075–80.
[19] Consonni D, Bertazzi PA. Health significance and statistical uncertainty. The value of P-value. Med Lav 2017;108:327–31.
[20] Laccourreye O, Maisonneuve H. Let's not twist and "spin-doctor" our articles. Eur Ann Otorhinolaryngol Head Neck Dis 2018;135:89.