Eur J Epidemiol (2017) 32:21–29
DOI 10.1007/s10654-016-0211-1

Statistical inference in abstracts of major medical and epidemiology journals 1975–2014: a systematic review

Andreas Stang[1,2] · Markus Deckert[1] · Charles Poole[3] · Kenneth J. Rothman[4]

Received: 5 September 2016 / Accepted: 7 November 2016 / Published online: 17 November 2016

Abstract Since its introduction in the twentieth century, null hypothesis significance testing (NHST), a hybrid of significance testing (ST) advocated by Fisher and null hypothesis testing (NHT) developed by Neyman and Pearson, has become widely adopted but has also been a source of debate. The principal alternative to such testing is estimation with point estimates and confidence intervals (CI). Our aim was to estimate time trends in NHST, ST, NHT and CI reporting in abstracts of major medical and epidemiological journals. We reviewed 89,533 abstracts in five major medical journals and seven major epidemiological journals, 1975–2014, and estimated time trends in the proportions of abstracts containing statistical inference. In those abstracts, we estimated time trends in the proportions relying on NHST and its major variants, ST and NHT, and in the proportions reporting CIs without explicit use of NHST (CI-only approach). The CI-only approach rose monotonically during the study period in the abstracts of all journals. In Epidemiology abstracts, as a result of the journal's editorial policy, the CI-only approach has always been the most common approach. In the other 11 journals, the NHST approach started out more common, but by 2014, this disparity had narrowed, disappeared or reversed in 9 of them. The exceptions were JAMA, New England Journal of Medicine, and Lancet abstracts, where the predominance of the NHST approach prevailed over time. In 2014, the CI-only approach is as popular as the NHST approach in the abstracts of 4 of the epidemiology journals: the American Journal of Epidemiology (48%), the Annals of Epidemiology (55%), Epidemiology (79%) and the International Journal of Epidemiology (52%). The reporting of CIs without explicitly interpreting them as statistical tests is becoming more common in abstracts, particularly in epidemiology journals. Although NHST is becoming less popular in abstracts of most epidemiology journals studied and some widely read medical journals, it is still very common in the abstracts of other widely read medical journals, especially in the hybrid form of ST and NHT in which *p* values are reported numerically along with declarations of the presence or absence of statistical significance.

Keywords Statistics · Confidence intervals · Statistics and numerical data

Introduction

Since its introduction early in the twentieth century, null hypothesis significance testing (NHST) has caused debate. It is a constantly mutating hybrid of Fisher significance testing (ST), in which *p* values are interpreted as continuous measures of evidence against hypotheses, and Neyman-Pearson

Electronic supplementary material The online version of this article (doi:10.1007/s10654-016-0211-1) contains supplementary material, which is available to authorized users.

✉ Andreas Stang
andreas.stang@uk-essen.de

[1] Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Hufelandstr. 55, 45147 Essen, Germany

[2] Department of Epidemiology, Boston University School of Public Health, 715 Albany St, Boston, MA 02118, USA

[3] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, CB #7435, 135 Dauer Drive, Chapel Hill, NC 27599-7435, USA

[4] RTI Health Solutions, Research Triangle Park, 200 Park Offices Dr, Durham, NC 27703, USA

null hypothesis testing (NHT), in which dichotomous categorizations of p values are used to accept or reject hypotheses [1]. A count in 2000 of over 300 warnings of limitations of ST, NHT and NHST [2] was followed a year later by a list of 402 references (http://warnercnr.colostate.edu/~anderson/thompson1.html, accessed Sept 30, 2015), among which we found 89 in biomedical publications. Despite the many cautions, NHST remains one of the most prevalent statistical procedures in the biomedical literature.

Experts, including many editors of biomedical journals, have repeatedly recommended de-emphasis of all variants of NHST in favour of estimation, in the form of point estimates and confidence intervals (CIs) [3–5]. Within the NHST context, guidance has strongly favored ST over NHT [6–10]. Although Walter wrote in 1995 that "the debate [related to the use of NHST] will undoubtedly continue for some time," [11] it might be news to some researchers that there is a debate. As it has continued, there are some indications that it has begun to create a movement away from strict adherence to NHT, if not to ST as well. For instance, in the Matrixx decision in 2011, the US Supreme Court unanimously ruled that admissible evidence of causality does not have to be statistically significant [12]. In 2015, the editors of a psychology journal banned not only NHT and ST, but CIs as well [13, 14].

Recently, the American Statistical Association (ASA) released a policy statement on statistical significance and p values including: "The widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process." [10] Despite this evidence, there is little documentation of the prevalence of NHT, ST and CIs in biomedical journals.

The few systematic reviews on the use of NHST and CIs in the biomedical and psychological literature that are available tend to focus on single journals [15–17]. Reviews across journals are scant [18–21]. In biomedical research in general, Chavalarias et al. recently used an automated text-mining analysis to extract data on p value reporting among 12 million MEDLINE abstracts and in more than 800,000 full-text articles in PubMedCentral (PMC) from 1990 through 2015. The proportion of abstracts containing p values (ST) or comparisons of p values with criterion values (NHT) increased from 7% in 1990 to 15% in 2014. In a subgroup analysis of a random sample 1000 abstracts from 1990 through 2015, only 2% reported CIs [19].

The aim of our systematic review is to take advantage of the capabilities of Medline's successor, PubMed, to investigate the presence of statistical inference in the form of NHT, ST and CIs in abstracts of major medical and epidemiology journals, especially the trends over time.

Materials and methods

Search and retrieval of abstracts

We searched PubMed (http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.PubMed_Coverage, accessed June 17–18, 2015) for all journal articles with abstracts published in 1975-2014 in the five medical journals (*Annals of Internal Medicine* [AIM], *British Medical Journal* [BMJ], *Journal of the American Medical Association* [JAMA], *Lancet*, *New England Journal of Medicine* [NEJM], and seven epidemiology journals (*American Journal of Epidemiology* [AJE], *Annals of Epidemiology* [ANE], *Epidemiology* [EPI], *European Journal of Epidemiology* [EJE], *International Journal of Epidemiology* [IJE], *Journal of Clinical Epidemiology* [JCE] (and its predecessor *Journal of Chronic Diseases*), *Journal of Epidemiology and Community Health* [JECH]) with the highest impact factors in 2014.

Abstract classifications

We used the index function of SAS to search for confidence intervals, p values, comparisons of p values with criterion values and the letter string "signif". We classified abstracts as follows: "Significance testing only" (ST-only) consists of reporting numerical p values (e.g., "$p = 0.02$") with neither comparisons of p values with thresholds nor significance terminology, regardless of CI reporting. "Null hypothesis testing only" (NHT-only) consists of comparing p values with thresholds (e.g., "$p \leq 0.05$") or use of significance terminology but no numerical p value reporting, regardless of CI reporting. "Null hypotheses significance testing" (NHST) includes any use of ST or NHT, regardless of CI reporting. "Any-CI" consists of any reporting of CIs, regardless of NHST reporting. Finally, "confidence interval only" (CI-only) consists of reporting of CIs without any NHST reporting.

Statistical methods

We estimated time trends by weighted nonparametric local regression smoothing (LOESS) [22, 23] with cubic local polynomials and a smoothing parameter of 0.5, which means that 50% of the data in each local neighborhood is used for the smoothing procedure. We derived weights by the score method, which incorporates a continuity correction for each annual proportion in each journal, and used these weights for the weighted LOESS fitted trend estimates [24]. For the visual display of smoothed time trends in the prevalence of abstract categories, we used stacked area charts [25].

Validation subsamples

We drew two stratified random samples of 400 abstracts (200 medical, 200 epidemiological). In the first, confined to ST-only abstracts with the letter string "signif," we determined the percentage in which the significance terminology was used in a nonstatistical sense. In the second, restricted to CI-only abstracts, we measured the percentage in which at least one CI for an estimated measure of association included the null value. As a third validation step, we compared our search algorithm-based approach of abstract classification with the thorough review of abstracts and full articles of the *American Journal of Epidemiology* by Savitz et al. [15], for NHST reporting by topic area (cancer, infectious disease and cardiovascular disease) in 1970, 1980, and 1990.

Results

Overall

Our review contained nearly 90,000 abstracts. The annual number of abstracts varied from 17 in the *Annals of Epidemiology* in 1990 to 722 in the *Lancet* in 2000. The annual number of abstracts containing statistical inference varied from 5 in the *Annals of Epidemiology* in 1990 and the *International Journal of Epidemiology* in 1978 to 310 in the *Lancet* in 2000 (Table 1).

Time trends in statistical inference in all abstracts

Figure 1a provides guidance for interpreting the time trends shown in Fig. 2 in the prevalence of statistical inference in all abstracts. The trend was an increase over time in every journal but the *Journal of Clinical Epidemiology*, where the prevalence remained steady at 40–50%. In most journals, the increase leveled off or began to decline in the late 1990s to early 2000s. In the *Journal of Epidemiology and Community Health*, the increase picked up again in the most recent decade. In three journals (*European Journal of Epidemiology, BMJ, JAMA*), the increase was steady and most dramatic in *JAMA*, from about 10% to about 90% over the four decades of the study period.

Time trends in NHST, its subtypes, and CIs in abstracts containing statistical inference

Figure 1b provides help in interpreting the time trends shown in Fig. 3 in the reporting of NHST, its variants (ST and NHT) and CIs in abstracts containing statistical inference. In *Epidemiology*, the prevalence of NHST in abstracts has remained low and the prevalence of CIs has

Table 1 Distribution of number of abstracts per calendar year and journal included in the systematic review of the publication years 1975 through 2014

	Calendar years	All Abstracts	Abstracts per calendar year			
			Overall		Containing any statistical inference	
			Min	Max	Min	Max
Medical journals		58,926	1058	1745	398	938
Ann Intern Med	1975–2014	8246	157	271	53	138
BMJ	1975–2014	12,149	144	463	87	221
JAMA	1975–2014	12,401	221	373	35	260
Lancet	1975–2014	16,750	266	722	107	310
N Engl J Med	1975–2014	9380	183	282	73	194
Epidemiology journals		30,607	138	1183	36	724
Am J Epidemiol	1975–2014	9239	98	397	26	265
Ann Epidemiol	1990–2014	2306	17	165	5	104
Epidemiology	1990–2014	2389	64	123	14	71
Eur J Epidemiol[a]	1985–2014	3124	46	167	10	98
Int J Epidemiol	1975–2014	4720	34	184	5	112
J Clinical Epidemiol[b]	1982–2014	4776	74	202	27	87
J Epidemiol & Community Health	1978–2014	4053	43	269	10	166

[a] In 2002, Medline did not contain abstracts of the *European Journal of Epidemiology*

[b] From 1982 through 1987, the predecessor *Journal of Chronic Diseases* was used

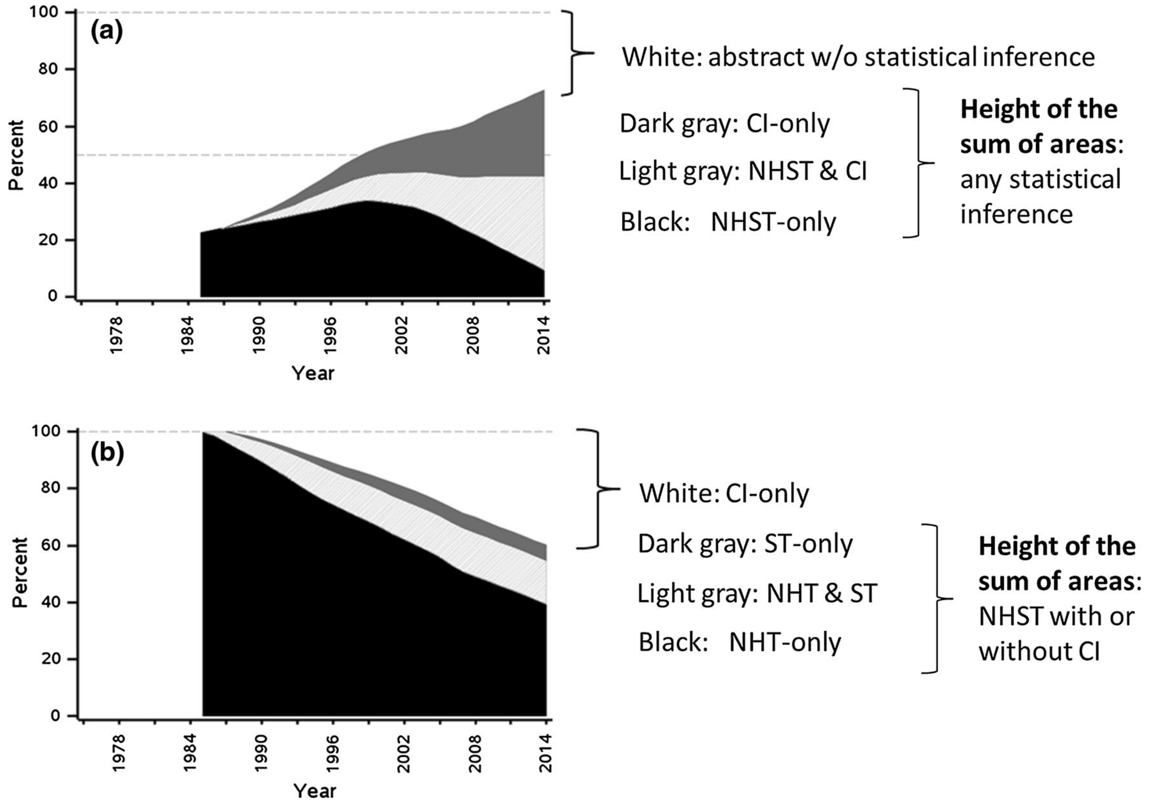

Fig. 1 Guidance for the interpretation of time trends in stacked area plots with the example of the European Journal of Epidemiology. **a** All abstracts (see Fig. 2). **b** Abstracts containing statistical inference (see Fig. 3). CI-only—reporting of confidence intervals without continuous or categorical *p* values and without significance terminology; NHST—null hypotheses significance testing includes any use of ST or NHT regardless of CI reporting; ST-only—significance testing only includes reporting of "p equals" with neither *p* value thresholds nor significance terminology regardless of CI reporting; NHT-only—null hypothesis testing only includes reporting of *p* value thresholds or significance terminology but no "p equals" reporting regardless of CI reporting; the white area from 1975 through 1984 indicates that the journal did not exist at that time period

remained high since the journal's inception in 1990. In the abstracts of every other journal, NHST has declined in prevalence and CIs have risen. The declines from an initial prevalence of approximately 100% in NHST have been most pronounced in the *International Journal of Epidemiology*, down to about 40%, and less so in the *American Journal of Epidemiology*, the *European Journal of Epidemiology*, the *Annals of Internal Medicine* and *BMJ* down to about 60%. The prevalence of any CIs rose from approximately zero to 80% or higher in the abstracts of seven journals (the *American Journal of Epidemiology*, the *European Journal of Epidemiology*, the *International Journal of Epidemiology*, the *Annals of Internal Medicine*, *BMJ*, *JAMA* and *Lancet*). In 2014, the CI-only approach is as popular as the NHST approach in the abstracts of 4 of the epidemiology journals: the *American Journal of Epidemiology* (48%), the *Annals of Epidemiology* (55%), *Epidemiology* (79%) and the *International Journal of Epidemiology* (52%).

Most recent period 2010–2014

The style of reporting statistical inference differed by journal category and among journals. In the most recent 5 years of our review (2010–2014), the prevalence of statistical inference in abstracts was lower in epidemiology journals (59%) than in medical journals (69%). Among abstracts containing statistical inference, the prevalence of CIs was lower in epidemiology than medical journals (74 vs 83%). However, the prevalence of CIs as the only means of statistical inference was higher in epidemiology journals than in medical journals (43 vs 22%). *Epidemiology* had the highest percentage of CI-only abstracts (84%), followed by the *International Journal of Epidemiology* (56%). All remaining journals had CI-only prevalences below 50%. In *JAMA, Lancet* and *New England Journal of Medicine*, it was under 25%. Also the style of reporting statistical test results differed by journal category and among journals. Medical journals more frequently presented a combination

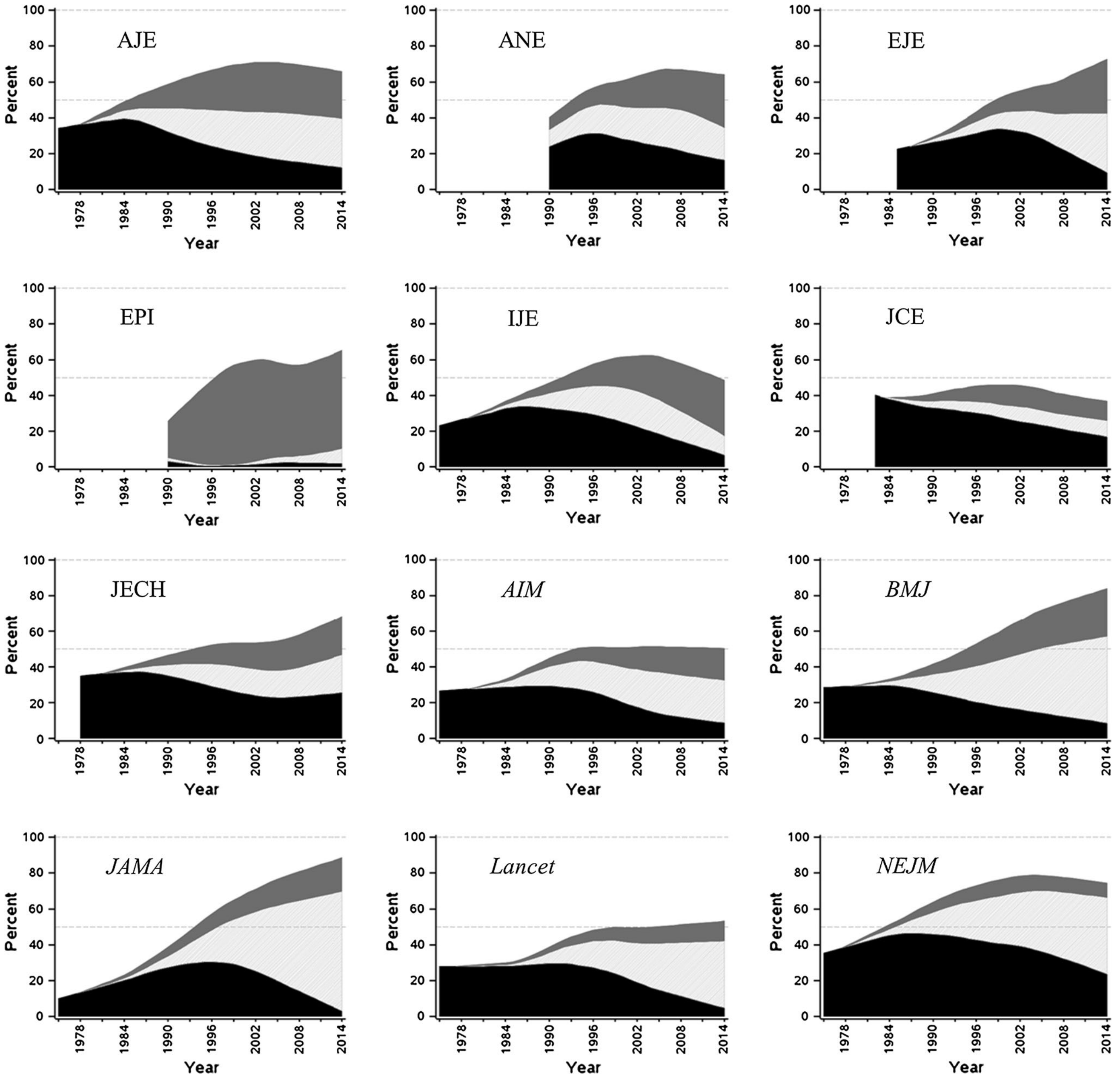

Fig. 2 Flexibly estimate time trends 1975–2004 in the prevalence of null hypothesis significance testing only, null hypothesis significance testing in combination with confidence intervals, and confidence intervals only in the abstracts of seven major epidemiology and five major medical journals. Flexibly (LOESS) fitted trend of the prevalence of statistical inference in abstracts; *black area* NHST-only; *light gray area* NHST combined with CIs; *dark gray area* CI-only; *white top area* percentage of abstracts that do not contain statistical inference; *AIM* Annals of Internal Medicine; *AJE* American Journal of Epidemiology; *ANE* Annals of Epidemiology; *EJE* European Journal of Epidemiology; *EPI* Epidemiology; *IJE* International Journal of Epidemiology; *JCE* Journal of Clinical Epidemiology; *JECH* Journal of Epidemiology and Community Health

of ST and NHT compared with epidemiology journals (Table 2).

Reviews of random samples of abstracts

In the 400 abstracts containing significance terminology without any other explicit NHST 89% (epidemiology 92%, medical 86%) clearly used that terminology in the statistical sense. In 1% (epidemiology 1%, medical 1%) we could not determine whether the significance language was statistical or substantive.

In the 400 abstracts that presented CIs without any NHST 20% (epidemiology 14%, medical journals 25%) were on topics or estimated measures not typically accompanied by NHST

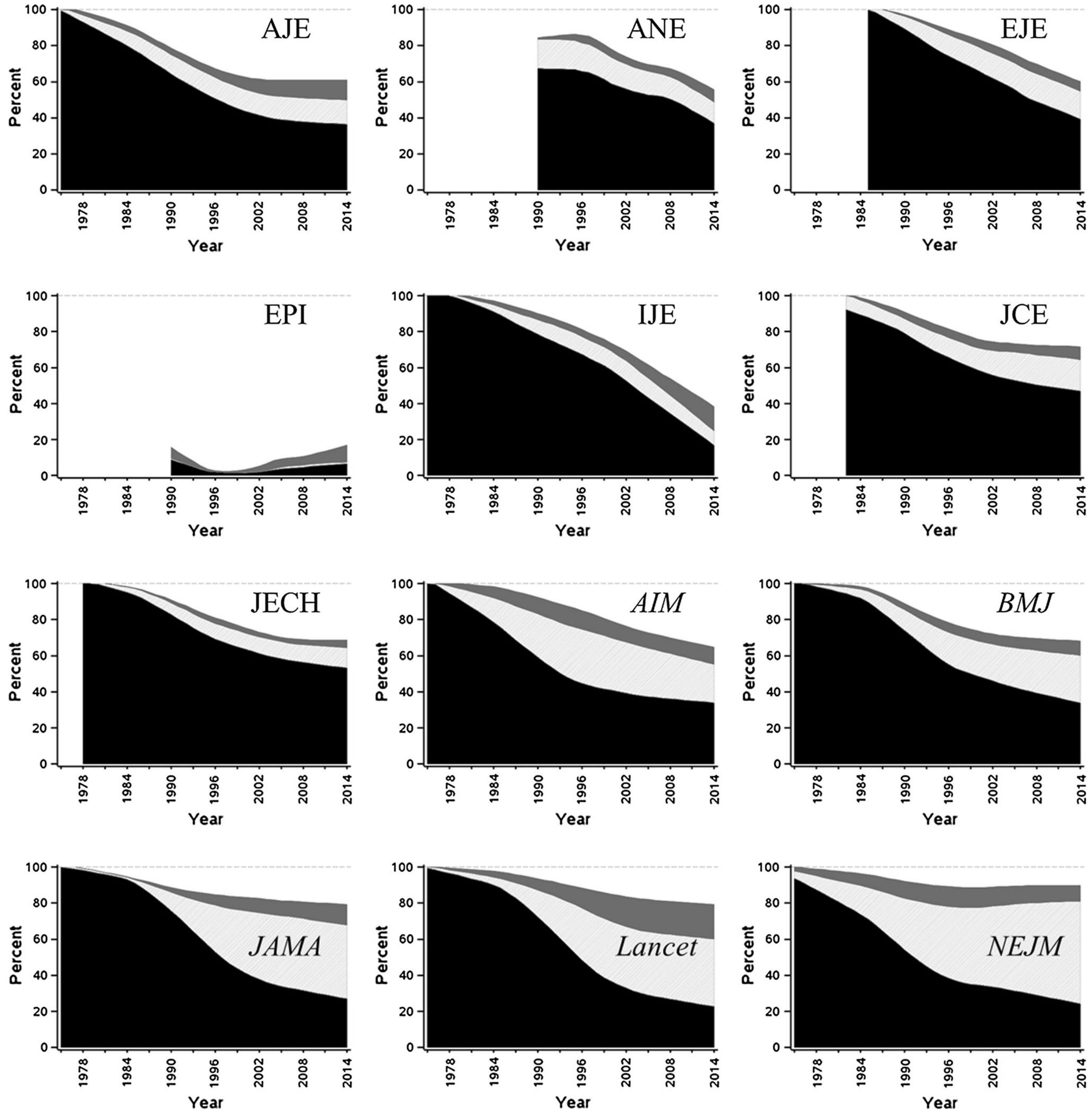

Fig. 3 Flexibly estimate time trends 1975–2004 in the prevalence of null hypothesis significance testing only, null hypothesis significance testing in combination with confidence intervals, and confidence intervals only in the abstracts of seven major epidemiology and five major medical journals containing any statistical inference. Flexibly (LOESS) fitted trend of the prevalence of statistical inference in abstracts; *black area* NHT-only; *light gray area* NHT & ST; *dark gray area* ST-only; *white top area* CI-only; *AIM* Annals of Internal Medicine; *AJE* American Journal of Epidemiology; *ANE* Annals of Epidemiology; *EJE* European Journal of Epidemiology; *EPI* Epidemiology; *IJE* International Journal of Epidemiology; *JCE* Journal of Clinical Epidemiology; *JECH* Journal of Epidemiology and Community Health

(disease frequency measures, diagnostic indices, measures of central tendency, methodological articles, etc.); among abstracts containing CIs for estimates of measures with null values, 59% reported only CIs that excluded the null value (epidemiology: 52%, medical: 68%); 31% reported at least one CI that included a null value and at least one that excluded it (epidemiology: 36%, medical: 25%), and 10% presented only CIs that included null values (epidemiology: 13%, medical: 7%).

Table 2 Prevalences of reporting of statistical inference in abstracts of the publication years 2010-2014

Journal	Total (n)	Any statistical inference (n)	Percent	Percentages among abstracts containing statistical inference (%)					
				Any CI	CI-only	Any NHST	ST-only	NHT-only	ST & NHT
All epidemiology journals	5339	3168	59	74	43	57	8	38	12
Am J Epidemiol	1531	1006	66	79	39	61	10	37	14
Ann Epidemiol	639	415	65	71	40	60	6	43	11
Epidemiology	488	301	62	95	84	16	7	7	1
Eur J Epidemiol	442	297	67	79	38	62	6	41	15
Int J Epidemiol	674	342	51	81	56	44	11	24	8
J Clin Epidemiol	783	301	38	52	29	71	7	48	16
J Epidemiol & Community Health	782	506	65	62	32	68	4	54	10

Journal	Total (n)	Any statistical inference (n)	Percent	Percentages (%)					
				Any CI	CI-only	Any NHST	ST-only	NHT-only	ST & NHT
All medical journals	5821	4022	69	83	22	78	11	30	37
Ann Int Med	842	411	49	79	34	66	14	54	31
BMJ	1160	942	81	88	30	70	8	35	27
JAMA	1160	995	86	92	21	79	11	30	38
Lancet	1452	763	53	86	20	79	19	25	36
N Engl J Med	1207	911	75	65	11	89	9	27	54

CI—reporting of any CI, CI-only—reporting of CIs without continuous or categorical *p* values and without significance terminology; NHST—null hypotheses significance testing includes any use of ST or NHT regardless of CI reporting; ST-only—significance testing only includes reporting of "*p* equals" with neither *p* value thresholds nor significance terminology regardless of CI reporting; NHT-only—null hypothesis testing only includes reporting of *p* value thresholds or significance terminology but no "*p* equals" reporting regardless of CI reporting

Comparison with a previous review

Our search algorithm-based approach revealed very similar percentages of abstracts that contained the NHST approach in the *American Journal of Epidemiology* as found in the thorough review by Savitz et al. [15] (Supplementary Table S3).

Discussion

The percentage of abstracts with statistical inference increased considerably from the mid 1970s to the most recent period among the high-impact medical journals and epidemiology journals in our review. This increase mirrors the general increase of statistical methods in biomedical journals. For example, in a review of statistical methods of full papers in the *New England Journal of Medicine*, Horton found that the percentage of articles that do not contain statistical methods or descriptive statistics only steadily decreased from 27% in 1978–1979 to 13% in 2004–2005 [26]. Similar time trends were found in other journals [27–29].

The prevalence of abstracts containing NHST without any CIs has dwinded from close to 100% in the journals that were published in the 1970s to below 25% today in every journal. CIs now appear more frequently than all forms of NHST combined in the abstracts of all but three journals: the *Journal of Clinical Epidemiology*, the *Journal of Epidemiology and Community Health* and the *New England Journal of Medicine*. The reporting of CIs without any explicit use of NHST has grown over time. The most impressive rises have been in the *International Journal of Epidemiology*, from zero in the mid 1970ies to about 60% today. CIs have always been the predominant mode of statistical inference and NHST in all its forms has been exceedingly rare in the abstracts of *Epidemiology*, since that journal's founding in 1990.

We believe that the steady stream of authoritative critiques promoting estimation over testing and, within testing, promoting ST over NHT has had a profound effect. For example, in 1988, more than 300 biomedical journals agreed to adhere to the manuscript guideline of the International Committee of Medical Journal Editors (ICMJE) that encouraged to present results with appropriate indicators of uncertainty (such as confidence intervals) [4]. In addition, the continuously growing number of articles and books that warn against NHST may have influenced authors and reviewers of journals over time.

Medical journals more frequently publish articles related to results of randomized studies than epidemiology journals. For drug approval, statistically significant results are usually requested by the Federal Drug Administration (FDA) [30]. Thus, this FDA request might have contributed to the higher percentage of abstracts containing the NHST approach in medical than epidemiology journals. One might ask, however, whether these journals are adhering to the dictates of regulatory authorities or whether the regulators are following what appears to be best scientific and statistical practice on display in the journal's pages. Clearly, there is a strong and growing consensus in the scientific and statistical community to prefer estimation over testing and that, within testing, to prefer ST over NHT. When systematic reviewers find literatures consisting only of statistical tests, whether in NHT or ST form, they know they will learn relatively little from those literatures. When literatures are filled with estimates of meaningful quantities, such as effect measures and CIs, much more will be learned [31–34].

Although our results for abstracts in the *American Journal of Epidemiology* in 1970, 1980, and 1990 were very similar to those of Savitz et al. [15], their review was enhanced by a close reading of the full articles. They found that by 1990, the "most common practice was to provide confidence intervals in results tables and to emphasize statistical significance tests in results texts" in that journal [15]. Furthermore, reporting practices differed by sections of the articles, with a lower percentage of CI reporting in abstracts than results tables [15]. Hence, practices, policies and trends in reporting in abstracts may not represent corresponding trends in full articles. Nonetheless, abstracts are of interest in their own right, as the abstract is often the part of a paper that is read first and sometimes the only part.

Our study is subject to some error. We measured only the overt statistical inference practices of authors in the abstracts of their papers. It is possible to engage in any of the testing practices covertly, especially NHT, without ever referring to whether or not hypotheses have been rejected and without using the words such as "significant," "significance" or "significantly" in their statistical sense. One form of covert NHST is to interpret every estimate CI that excludes the null value as an association and every estimate that includes the null value as no association [9]. We were unable to determine the frequency of this and similar practices, as it would have required us to read full articles even more closely than Savitz et al. [15] did. We suspect that at least some of the papers whose abstracts we classified as "CI-only" actually engaged in NHST and that at least some of the papers whose abstracts we classified as "ST-only" actually engaged in NHT.

Although 41% of the CIs for measures with null values in our random sample included that null value, but we did not attempt to discern whether the authors interpreted those CIs as associations or as no association. In addition, the 10% of abstracts that clearly used significance terminology only in a nonstatistical sense is unmistakable evidence of classification error on our part with regard to abstracts, though some of the authors may have engaged in overt or covert NHST nonetheless in the full articles.

Acknowledgements The authors thank Sander Greenland, DrPH for valuable comments on an early draft.

Funding Andreas Stang receives a grant from the German Federal Ministry of Education and Science (BMBF), Grant Number 01ER1305.

Compliance with ethical standards

Conflict of interest None of the authors declares a conflict of interest.

Authors' contributions AS, MD, CP, and KJR were involved in the study design. AS and MD performed the statistical analyses. AS wrote the first draft of the report. All authors contributed to the final version.

References

1. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. The empire of chance. How probability changed science and everyday life. Cambridge: Cambridge University Press; 1989.
2. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manag. 2000;64(4):912–23.
3. Rothman KJ. A show of confidence. N Engl J Med. 1978;299(24):1362–3.
4. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Br Med J (Clin Res Ed). 1988;296(6619):401–5.
5. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986;292(6522):746–50.
6. Rothman KJ. Significance questing. Ann Intern Med. 1986;105(3):445–7.
7. Weinberg CR. It's time to rehabilitate the P-value. Epidemiology. 2001;12(3):288–90.
8. Sterne JA, Davey SG. Sifting the evidence-what's wrong with significance tests? BMJ. 2001;322(7280):226–31.
9. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50.
10. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129–33.
11. Walter SD. Methods of reporting statistical results from medical research studies. Am J Epidemiol. 1995;141(10):896–906.
12. Gastwirth JL. Statistical considerations support the supreme court's decision in Matrixx Initiatives v. Siracusano. Jurimetrics. 2012;52:155–75.
13. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych. 2015;37:1–2.
14. Anonymous. Psychology journal bans P values. Nature 2015; 519:9.
15. Savitz DA, Tolo KA, Poole C. Statistical significance testing in the American Journal of Epidemiology, 1970–1990. Am J Epidemiol. 1994;139(10):1047–52.

16. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol Sci. 2004;15(2):119–26.
17. MacArthur RD, Jackson GG. An evaluation of the use of statistical methodology in the. J Infect Dis. 1984;149(3):349–54.
18. Vacha-Haase T, Nilsson JE, Reetz DR, Lance TS, Thompson B. Reporting practices and APA editorial policies regarding statistical significance and effect size. Theory Psychol. 2000;10(3):413–25.
19. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990–2015. JAMA. 2016;315(11):1141–8.
20. Fritz A, Scherndl T, Kühlberger A. A comprehensive review of reporting practices in psychological journals: are effect sizes really enough? Theory Psychol. 2012;23(1):98–112.
21. Thompson B. Journal editorial policies regarding statistical significance tests: heat is to fire as p is to importance. Educ Psychol Rev. 1999;11(2):157–69.
22. Cleveland WS, Devlin S, Grosse E. Regression by local fitting. J Econom. 1988;37:87–114.
23. Cleveland WS, Grosse E. Computational methods for local regression. Stat Comput. 1991;1:47–62.
24. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17(8):857–72.
25. Milne PH. Presentation graphics for engineering, science, and business. London: E & FN Spon; 2005.
26. Horton NJ, Switzer SS. Statistical methods in the journal. N Engl J Med. 2005;353(18):1977–9.
27. Felson DT, Cupples LA, Meenan RF. Misuse of statistical methods in Arthritis and Rheumatism. 1982 versus 1967-68. Arthritis Rheum. 1984;27(9):1018–22.
28. Arnold LD, Braganza M, Salih R, Colditz GA. Statistical trends in the Journal of the American Medical Association and implications for training across the continuum of medical education. PLoS ONE. 2013;8(10):e77301.
29. Jin Z, Yu D, Zhang L, et al. A retrospective survey of research design and statistical analyses in selected Chinese medical journals in 1998 and 2008. PLoS ONE. 2010;5(5):e10822.
30. Guidance for Industry. E9 Statistical Principles for Clinical Trials. Food and Drug Administration 1998. www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf. Accessed Oct 4, 2015.
31. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 510 (updated March 2011): Cochrane Collaboration (www.handbook.cochrane.com); 2011.
32. Koricheva J, Gurevitch J. Place of meta-analysis among other methods of research synthesis. In: Koricheva J, Gurevitch J, Mengerson K, editors. Handbook of meta-analysis in ecology and evolution. Princeton: Princeton University Press; 2013. p. 1–13.
33. Freemantle N, Geddes J. Understanding and interpreting systematic reviews and meta-analyses. Part 2: meta-analyses. *Evid Based*. Mental Health. 1998;1:102–4.
34. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2009. P. 251–5, 297–302, 325–31.