The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

The Use of Statistics in Medical Research

A Comparison of *The New England Journal of Medicine* and *Nature Medicine*

Alexander M Strasak, Qamruz Zaman, Gerhard Marinell, Karl P Pfeiffer & Hanno Ulmer

To cite this article: Alexander M Strasak, Qamruz Zaman, Gerhard Marinell, Karl P Pfeiffer & Hanno Ulmer (2007) The Use of Statistics in Medical Research, The American Statistician, 61:1, 47-55, DOI: 10.1198/000313007X170242

To link to this article: https://doi.org/10.1198/000313007X170242

Published online: 01 Jan 2012.

MEDICINE

The Use of Statistics in Medical Research: A Comparison of *The New England Journal of Medicine* and *Nature Medicine*

Alexander M. STRASAK, Qamruz ZAMAN, Gerhard MARINELL, Karl P. PFEIFFER, and Hanno ULMER

There is widespread evidence of the extensive use of statistical methods in medical research. Just the same, standards are generally low and a growing body of literature points to statistical errors in most medical journals. However, there is no comprehensive study contrasting the top medical journals of basic and clinical science for recent practice in their use of statistics.

All original research articles in Volume 10, Numbers 1–6 of *Nature Medicine* (*Nat Med*) and Volume 350, Numbers 1–26 of *The New England Journal of Medicine* (*NEJM*) were screened for their statistical content. Types, frequencies, and complexity of applied statistical methods were systematically recorded. A 46-item checklist was used to evaluate statistical quality for a subgroup of papers.

94.5 percent (95% CI 87.6–98.2) of *NEJM* articles and 82.4 percent (95% CI 65.5–93.2) of *Nat Med* articles contained inferential statistics. *NEJM* papers were significantly more likely to use advanced statistical methods ($p < 0.0001$). Statistical errors were identified in a considerable proportion of articles, although not always serious in nature. Documentation of applied statistical methods was generally poor and insufficient, particularly in *Nat Med*.

Compared to 1983, a vast increase in usage and complexity of statistical methods could be observed for *NEJM* papers. This does not necessarily hold true for *Nat Med* papers, as the results of the study indicate that basic science sticks with basic analysis. As statistical errors seem to remain common in medical literature, closer attention to statistical methodology should be seriously considered to raise standards.

KEY WORDS: Complexity; Errors and shortcomings; Statistical methods in medical journals; Techniques.

Alexander M. Strasak is Statistician, Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Austria (E-mail: *alexander.strasak@i-med.ac.at*). Qamruz Zaman is Ph.D. Student, Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Austria (E-mail: *qamruz.zaman@uibk.ac.at*). Gerhard Marinell is Statistician, Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck, Austria (E-mail: *gerhard.marinell@uibk.ac.at*). Karl P. Pfeiffer is Statistician, Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Austria (E-mail: *karl-peter.pfeiffer@i-med.ac.at*). Hanno Ulmer is Statistician, Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Austria (E-mail: *hanno.ulmer@i-med.ac.at*).

1. INTRODUCTION

Over the past decades, a great increase in the use of statistical methods has been documented for a wide range of medical journals (Altman 1982, 1991, 2000). Favored by the availability of multifaceted statistical software packages, a trend toward the usage of more sophisticated techniques can be observed. Nevertheless there is also strong evidence that, in particular, simple methods such as *t* tests or chi-square tests remain in common usage (Emerson and Colditz 1983; Colditz and Emerson 1985; Menegazzi et al. 1991; Cardiel and Goldsmith 1995; Huang, LaBerge, Lu, and Glidden 2002; Reed III, Salen, and Bagher 2003).

The use of statistics in medical journals has been subjected to considerable debate in recent years and there is wide consensus that standards are generally low, as a high proportion of published medical research contains statistical errors (Schor and Karten 1966; Gore, Jones, and Rytter 1977; MacArthur and Jackson 1984; Pocock, Hughes, and Lee 1987; McKinney et al. 1989; Gardner and Bond 1990; Kanter and Taylor 1994; Porter 1999; Cooper, Shriger, and Close 2002; García-Berthou and Alcaraz 2004). The misuse of statistics in medical research has therefore been widely discussed, and it has been pointed out that it is both unethical and can have serious clinical consequences (Altman 1981; Gardenier and Resnik 2002). As a result, there was respectable effort from many medical journals to enhance quality of statistics by adopting statistical guidelines and recommendations for authors or by sharpening the statistical review of incoming manuscripts (Altman et al. 1983; Murray 1991; Gore, Jones, and Thompson 1992; Goodman, Altman, and George 1998; Altman 1998; Moher, Schulz, and Altman 2001; Barron 2006). Nonetheless also very recent studies, although generally focused on specific statistical details, point toward major problems (Cooper et al. 2002; Olsen 2003; García-Berthou and Alcaraz 2004; Marshall 2004).

In the present study we report current practices regarding the use of statistics in medicine by contrasting the top journal for basic science *Nature Medicine* [impact factor = 28.88 (Journal Citation Report 2005, Institute of Scientific Information, Thomson Corp.); ranked no. 1 in area "medicine, research and experimental"] with the top journal for clinical science *The New England Journal of Medicine* [impact factor = 44.02 (Journal Citation Report 2005, Institute of Scientific Information, Thomson Corp.); ranked no. 1 in area "medicine, general, and internal"].

 DOI: 10.1198/000313007X170242

Nature Medicine is a biomedical research journal covering fields such as cancer biology, cardiovascular research, gene therapy, or immunology. Original research articles published in the journal range from basic findings that have clear implications for disease pathogenesis and therapy to the earliest phases of human investigations, whereby the journal aims to publish the most relevant research papers that form the foundations of tomorrow's medicine. On the other hand, *The New England Journal of Medicine* is the oldest continuously published medical journal in the world, with more than 500,000 regular readers in more than 177 countries. It aims to publish the very best information from research, at the interface of biomedical science, internal medicine, and clinical practice, in an understandable and clinically useful format.

The first aim of the study was to investigate and to compare the statistical methods used in research papers of both journals and to assess complexity of statistical analyses. The second was to evaluate the quantity and character of statistical misuse and statistical errors. Although the statistical content of several medical journals has been reviewed over the past decades (Altman 1981, 1982, 1991, 2000; Colditz and Emerson 1985; Menegazzi et al. 1991; Olsen 2003), there is no recent and comprehensive evaluation comparing the statistical methods used in a top clinical journal to those used in a top basic science journal. Questions regarding their recent use therefore remain largely unanswered. The results of the study allow for the ongoing monitoring of possible trends in statistics usage as well as for an up-to-date in-depth comparison between clinical research literature which is closely linked to modern statistics and basic research which is emphatically not. The most frequent errors and omissions observed in a detailed quality assessment of both journals are outlined. A comprehensive 46-item checklist for the statistical evaluation of medical manuscripts is presented.

2. MATERIALS AND METHODS

All *original research articles* published during the first half of year 2004 in Volume 350, Numbers 1–26 of *The New England Journal of Medicine* and Volume 10, Numbers 1–6 of *Nature Medicine* were included for a bibliometric analysis. Editorials, letters, case reports, and review articles were excluded. There were a total of 91 papers for *The New England Journal of Medicine* (*NEJM*) and 34 papers for *Nature Medicine* (*Nat Med*).

All 125 papers were manually reviewed for their statistical content. Types and frequencies of applied statistical methods were systematically recorded and classified into 17 categories, earlier used by Emerson and Colditz (1983). Papers containing statistical analyses beyond descriptive statistics were classified into Basic Analyses or Advanced Analyses according to sophistication of applied statistical techniques. Thereby *t* tests, simple contingency table analysis (including chi-square- or Fisher's exact tests), nonparametric methods, one-way ANOVA, simple correlation and regression techniques were considered as Basic Analyses, whereas all papers employing any sophisticated statistical technique beyond those listed above were classified into the category Advanced Analyses. Papers containing any method of multivariate analysis (e.g. MANOVA, MANCOVA), statistical modeling, advanced contingency table analysis, epidemiologic statistics, or survival analysis obligatory were categorized as Advanced Analyses. For each paper the number of different statistical techniques used was counted in order to determine the amount of various statistical methods involved in every article.

A subgroup of 53 papers (*NEJM* = 31, *Nat Med* = 22) was further selected for a detailed quality assessment of statistical methods used. For this purpose, a standardized 46-item checklist was developed on the basis of manifold literature related to the topic (Gore et al. 1977; White 1979; Gardner et al. 1983; MacArthur and Jackson 1984; Felson, Cupples, and Meenan 1984; Thorn et al. 1985; Avram et al. 1985; Gardner, Machin, and Campbell 1986; Gardner and Bond 1990; Goodman and Hughes 1992; Dar, Serlin, and Omer 1994; Kanter and Taylor 1994; McGuigan 1995; McCance 1995; Kuzon, Urbanchek, and McCabe 1996; Welch II and Gabbe 1996; Altman 1998; Moher et al. 2001; Olsen 2003). Although this assessment tool clearly cannot cover aspects of all statistical methods incorporated in modern medical research, it is to our knowledge one of the most comprehensive lists presented yet. While predominately focusing on the issue of statistical significance testing, the checklist includes multifaceted statistical aspects of study design, statistical analysis, documentation of applied statistical methods, as well as presentation and interpretation of study findings (see Appendix, p. 54).

Assortment of papers for this evaluation was done according to predefined inclusion criteria, with insistence on (1) the use of inferential methods beyond descriptive statistics and (2) the use of one or more elementary statistical significance tests to evaluate the primary outcome measure(s) in a paper (e.g., *t* test, chi-square test, Fisher's exact test, Mann-Whitney-U-test, Wilcoxon test, etc.). Articles focusing on advanced statistical techniques and/or statistical modeling, which were not explicitly covered by the checklist, were excluded from further assessment.

After detailed and critical examination of all sections, tables, and figures, the standardized checklist was completed by the first author (A.M.S.) for each of the 53 papers. If an assessment of a single item was not clear or vague due to limits of tolerance typically associated with specific statistical issues, this item was independently reviewed by a second statistician (H.U.) and then assessed together. If a paper contained insufficient information to assess a specific item of the checklist, an evaluation "unable to assess/not clear" was mandatory. A classification "error committed" was restricted to unquestioned, obvious issues that could clearly be identified.

For the journals under investigation, two prespecified hypotheses regarding potential differences in the proportions of papers using (1) inferential methods and (2) advanced analyses were examined. Statistical analysis was done by two-tailed Fisher's exact tests with a level of significance set at 0.05. Exact 95% confidence intervals were computed by the Clopper and Pearson (1934) method.

3. RESULTS

Table 1 show the types and frequencies of statistical methods in all original research papers of Volume 350, Numbers 1–26 in *NEJM* and Volume 10, Numbers 1–6 in *Nat Med.* Of 91 papers analyzed from *NEJM*, 94.5 percent (95% CI 87.6–

Table 1. Types, Frequencies, and Complexity of Statistics

Types and frequencies of statistical methods‡	New England Journal of Medicine (n = 91) n	%	Nature Medicine* (n = 34) n	%	P Value for comparison†
No statistical methods	2	2.2	1	2.9	
Descriptive statistics only	3	3.3	5	14.7	
Inferential methods	86	94.5	28	82.4	0.068
t tests	32	35.2	14	41.2	
Contingency table analysis					
Basic (χ^2-, Fishers exact test)	42	46.2	0	0.0	
Advanced	6	6.6	0	0.0	
Nonparametric tests	24	26.4	7	20.6	
Analysis of Variance					
Basic (one-way ANOVA)	6	6.6	9	26.5	
Advanced	6	6.6	1	2.9	
Correlation coefficients	12	13.2	2	5.9	
Regression					
Basic (simple-linear regression)	4	4.4	1	2.9	
Advanced	27	29.7	0	0.0	
Epidemiologic methods	25	27.5	0	0.0	
Survival Analysis	39	42.9	4	11.8	
Other methods	15	16.5	5	14.7	
Unidentified method/test	1	1.1	10	29.4	
Confidence intervals	61	67.0	0	0.0	
Complexity of statistical analyses					
No. of different inferential methods					
Only 1 method	15	16.5	10	29.4	
2 or 3 methods	39	42.9	15	44.1	
4 or 5 methods	22	24.2	3	8.8	
More than 5 methods	10	11.0	0	0.0	
No/descriptive/unidentified methods	6	6.6	12	35.3	
Basic analyses§	15	16.5	15	44.1	
Advanced analyses¶	70	76.9	7	20.6	<0.0001

* Six papers did not specify or document any of the applied statistical methods and therefore could not be used for further classification and assessment.

† *P* values determined by Fishers Exact test, 2-tailed.

‡ As many papers contained usage of more than one category of statistical methods listed, numbers presented do not add up to the whole of papers reviewed, respectively to 100.0 percent. A full explanation for the categories listed is given by Emerson/Colditz.

§ *t* tests, Contingency Table Analysis Basic, Nonparametric tests, ANOVA Basic, Correlation coefficients, Regression Basic.

¶ Contingency Table Analysis Advanced, ANOVA Advanced, Regression Advanced, Epidemiologic methods, Survival Analysis, other methods.
If application of even only one of these methods listed could be identified in a paper, classification "Advanced Analyses" was obligatory.

98.2) contained some kind of inferential statistics. The corresponding number for *Nat Med* papers adds up to 82.4 percent (95% CI 65.5–93.2). No statistically significant difference in the frequency of papers using inferential methods could be detected between the examined journals ($p = 0.068$). The most frequently used inferential statistics in *NEJM* were methods of survival analysis (Kaplan-Meier, Mantel-Cox, log rank test, life table analysis), which could be observed in 39 of 91 papers (42.9 percent), closely followed by simple *t* tests and chi-square tests with a frequency of 35.2 percent each. For *Nat Med* papers the most commonly reported inferential methods were *t* tests (41.2 percent) and analysis of variance (29.4 percent), whereas 10 papers (29.4 percent) included usage of "unidentified tests," as the authors failed to name the statistical procedures employed for generating the presented *p* values. Confidence intervals were entirely absent in *Nat Med* papers, but could be identified in 61 papers from *NEJM* (67.0 percent).

Complexity and sophistication of statistical analyses was significantly higher for papers from *NEJM* ($p < 0.0001$, Table 1). Only 16.5 percent of these papers were classified Basic Analyses for constricting statistical evaluation to exclusively elementary techniques like *t* tests, χ^2 tests, Fisher's exact tests, simple nonparametric tests, one-way ANOVA, or linear regression and correlation. 70 *NEJM* papers (76.9 percent, 95%CI 66.9–85.1) reported the use of at least one more sophisticated method beyond those listed above and therefore were classified Advanced Analyses. The corresponding number for *Nat Med* was 7 papers (20.6 percent, 95%CI 8.7–37.9).

Tables 2 and 3 show the types and frequencies of statistical errors and deficiencies, identified in a subsequent quality assessment of employed statistical methods. The most common shortcoming related to the design of a study was the failure to conduct and/or properly report statistical sample size estimation or power calculation. Although crucial to a modern and

*Table 2. Statistical Errors, Flaws, and Deficiencies related to the Design of a Study and Statistical Analysis**

Category	*New England Journal of Medicine* (n = 31)		*Nature Medicine** (n = 22)	
	n	%	*n*	%
Design of Study				
No sample size calculation/power calculation (overall)	13	41.9	22	100.0
Prospective study design	10	32.3	22	100.0
Retrospective study design	1	3.2	0	0.0
Study design not clearly classifiable†	2	6.5	0	0.0
Failure to use and report randomization when possible‡	1	3.2	13	59.1
Method of randomization/allocation to intervention not clearly stated‡	6	19.4	4	18.2
Data Analysis				
Use of a wrong or suboptimal statistical test	5	16.1	6	27.3
Incompatibility of statistical test with type of data examined	4	12.9	1	4.5
Inappropriate use of parametric methods	2	6.5	5	22.7
Use of an inappropriate test for the hypothesis under investigation	3	9.7	2	9.1
Failure to include a multiple-comparison correction/α-level correction	11	35.5	6	27.3
Special errors with Student's *t* test				
Failure to prove/report that test assumptionsare not violated	16	51.6	13	59.1§
Improper multiple pairwise comparisons without α-level correction	2	6.5	3	13.6
Special errors with χ^2 tests				
No Yates correction if small numbers	4	12.9	0	0.0¶
Use of chi-square when expected numbers in a cell are < 5	9	29.0	0	0.0

* Papers with checklist assessment "unable to assess/not clear" not shown.

† Papers did not contain sufficient information to clearly classify design of study.

‡ Due to different study designs, statistical assessment for this category was not possible for all papers.

§As only 13 papers reported usage of *t* tests, the actual error rate for failing to prove *t* test-assumptions equals 100.0 percent.

¶None of the *Nature Medicine* papers reviewed reported the use of χ^2 tests.

efficient design of a study, both issues were entirely absent in all 22 reviewed and prospectively designed *Nat Med* papers, as none contained any statement on a priori sample size calculation. However, this probably might be due to editorial policy of the journal. Five of 31 papers from *NEJM* (16.1 percent) contained usage of wrong or suboptimal statistical tests, either because of incompatibility of test with examined data, inappropriate use of parametric methods, or use of an inappropriate statistical test for the scientific hypothesis under investigation. The corresponding proportion for *Nat Med* papers equals 27.3 percent, and was therefore slightly higher. Other frequently observed omissions were inappropriate usage of standard errors of the mean (SEM) for describing variability of study data and inaccurate reporting of arbitrary thresholds (e.g. "$p < 0.05$", "$p = ns$") instead of specifying exactly obtained p values. Common statistical errors related to interpretation of study findings were disregard for Type II errors when reporting nonsignificant results and neglect of multiple testing problems, associated with multiple study endpoints, which were observed in 32.3 and 27.3 percent of *NEJM* and *Nat Med* papers, respectively.

Because of intense and persistent deficiencies in documentation of employed statistical methods, especially in *Nat Med* papers, it was in general fairly difficult to determine the accuracy and appropriateness of statistical analyses and interpretation of study results. Due to insufficient information and/or limits of tolerance especially the checklist items "use of a wrong or suboptimal statistical test" (including subitems) and "drawing conclusions not supported by the study data" were most difficult to assess. Regarding these two items, 24 papers with an initial checklist assessment "unable to assess/not clear" were independently re-reviewed by a second statistician (H.U.), but still a rate of nearly 50 percent remained unclear ("unable to assess/not clear").

4. DISCUSSION

The results of the present study give up-to-date evidence for the expanding use of inferential statistics in modern medical research. Compared with the early results of Emerson and Colditz (1983) for *The New England Journal of Medicine*, a vast increase in the application of inferential methods could be outlined, as the proportion of papers having analytical character more than doubled from 42.0 percent in 1983 to 94.5 percent in 2004. In particular, methods of Survival Analysis, which were virtually absent in 1983, are now in common use in this journal as also recently documented by Horton and Switzer (2005).

Although complexity of statistical analyses in *The New England Journal of Medicine* continued to increase since 1983 and was rather sophisticated in 2004, this does not necessarily hold for papers from *Nature Medicine*, which gives reason to hypothesize that basic science sticks with basic analyses. However, the majority of research reports published in *Nature Medicine* are based on animal studies, whereas *The New England Journal of Medicine* predominately publishes studies on human subjects.

*Table 3. Statistical Errors, Flaws, and Deficiencies related to Documentation, Presentation and Interpretation**

Category	*New England Journal of Medicine* (n = 31)		*Nature Medicine* (n = 22)	
	n	%	*n*	%
Documentation				
Failure to specify/define all applied statistical tests clearly and correctly	20	64.5	20	90.9
Failure to state number of tails	8	25.8	20	90.9
Failure to state if test was paired or unpaired	18	58.1	17	77.3
Failure to specify which test was performed on a given set of data†	12	8.7	7	31.8
"Where appropriate" statement	7	22.6	2	9.1
Failure to state which values of *p* indicate statistical significance	14	45.2	15	68.2
Presentation				
"Mean" but no indication of variability of data‡	2	6.5	0	0.0
Giving standard error (SE) instead of SD for statistical description	8	25.8	16	72.7
Failure to define $+/-$ notion for describing variability; use of unlabeled error bars	2	6.5	4	18.2
Use of arithmetic mean/SD to describe non-normal or ordinal data	4	12.9	0	0.0
No confidence intervals for main effect size measures presented	14	45.2	21	95.5
"p = NS," "$p < 0.05$," "$p > 0.05$" etc. instead of reporting exact *p* values	6	19.4	19	86.4
Interpretation§				
"non significant" treated/interpreted as "no effect"/"no difference"	1	3.2	0	0.0
Significance claimed without data analysis or statistical test mentioned	1	3.2	2	9.1
Disregard for Type II error when reporting nonsignificant results	5	16.1	5	22.7
Missing discussion of the problem of multiple significance testing if occurred	10	32.3	6	27.3

* Papers with checklist assessment "unable to assess/not clear" not shown.

† Statistical assessment for this category only was possible when more than one type of statistical test was reported and performed.

‡ "Mean" corresponds to any kind of statistical measure of central tendency.

§ Assessment of interpretation was exclusively restricted to statistically based conclusions of study findings. Correctness of conclusions toward medical relevance, clinical importance or future implications of a study were not evaluated in this analysis.

As a consequence, statistical demands are different between basic and clinical research. Typical characteristics for animal studies are the use of experimental designs including a high degree of invasiveness as well as less intraindividual variation due to usage of genetically identical species. This reduces the necessity for the application of multivariate analysis to adjust for possible confounding which is typically present in clinical settings. Other advanced methods such as survival analysis that were frequently observed in *The New England Journal of Medicine* probably are less likely to figure in *Nature Medicine* for the same reason. The smaller sample sizes associated with animal studies further lessen the possibility of applying sophisticated statistical techniques.

Another difference between the basic and clinical research journals reviewed, regards the scale of measurements towards a greater use of continuous data in *Nature Medicine*. Whereas almost every second article in *The New England Journal of Medicine* contained chi-square testing, there was no such test to be found in the sampled articles of *Nature Medicine* and it rather appears that Student's *t* test is still the most applied inferential method in basic science. However, clinical studies are more observational in nature and therefore are characterized by a greater use of ordinal and nominal data.

Referring to our results, it can be argued with caution, that the journal impact factor does not seem to be a meaningful predictor for the statistical quality of published medical research, as types and frequencies of statistical errors and flaws identified in this review, did not differ substantially from earlier results, found in similar studies for a wide range of other medical journals from various disciplines (Schor and Karten 1966; Gore et al. 1977; MacArthur and Jackson 1984; Pocock et al. 1987; McKinney et al. 1989; Gardner and Bond 1990; Kanter and Taylor 1994; Welch II and Gabbe 1996; Davies 1998; Olsen 2003; Marshall 2004). With the exception of manuscripts submitted to a few high-impact journals in clinical research, there is still a very small percentage of papers undergoing statistical reviewing and only a minority of medical journals has yet published their statistical guidelines for authors. Especially in basic science journals this trend seems to be at the very beginning although recently addressed in a *Nature Medicine* editorial, which was exclusively devoted to statistical "sloppiness" of research papers in the journal (Vol. 11, No. 1, p. 1). Concerning this, the editors also announced the release of statistical guidelines for authors in near future. Nevertheless, the results of the in-depth statistical quality assessment strongly suggest that a more clearly stated statistical policy, a more explicit set of instructions to authors, and closer editorial attention to statistical methodology, starting at the pre-publication phase of a manuscript, should seriously be considered not only by *Nature Medicine*. In this context, the 46-item checklist, presented in this study, could feasibly be used for a comprehensive statistical reviewing of a wide range of medical manuscripts, as it contains aspects of virtually all stages of scientific research.

In addition, the role and the involvement of statisticians seem

to be very different in basic and clinical research settings. Whereas clinicians tend to make frequent and in time use of statistical consulting, basic researchers often do not and are largely planning and conducting their research without the help of professional statistical advice. A greater involvement of statisticians serving in editorial as well as in reviewing tasks could therefore markedly improve quality of scientific research, especially in the area of basic science. The observed shortcomings in the design of several studies suggest that statisticians should be involved at an earlier stage of research and greater responsibilities should be taken by them regarding the planning phase of a study, as well as in data collection and data management issues.

On part of the authors of medical manuscripts it should be considered that a research report fails in its task of informing a reader if there is insufficient information for a critical assessment of its findings. Thus, as well as using adequate statistical methods, it is essential to describe the statistical methods employed with enough detail to enable a knowledgeable reader to recalculate important study findings. Unfortunately, this was not possible for more than 90% of papers from *Nature Medicine*, as the authors failed to properly specify and define all of the statistical tests used. More emphasis should be given to the magnitude of treatment differences and to statistical estimation techniques, than to solely rely on uncritical significance testing, as continuously observed in *Nature Medicine* papers.

Despite the fact that the majority of statistical errors and deficiencies identified in both journals were of minor importance and frequently related to insufficient documentation, some kind of "error continuation" could be observed, which means that single errors seemed to occur more often in single journals. Especially the erroneous indication of the SEM to describe variability of study data (Davies 1998) seems to be a common problem in *Nature Medicine*, as this statistical error nearly occurred in three out of four papers. Five other common statistical errors and shortcomings that were frequently observed in research papers from both journals reviewed are listed below:

- Use of a wrong or suboptimal statistical test
- Failure to adjust for multiple comparisons
- Failure to indicate confidence intervals for main effect size measures
- Failure to define all statistical tests used clear and correctly
- Failure to conduct a priori sample size estimation

Although in general those errors may not inevitably lead to incorrect conclusions in a paper, they still have considerable negative effect on the scientific impact of a research project by diminishing reliability, validity, and verifiability of study results. As the five listed errors can easily be avoided, particular attention should be given to them by both basic and clinical researchers.

There are mainly two limitations of our study concerning the quantitative and qualitative assessment of the articles reviewed. First, the proportion of papers selected from *The New England Journal of Medicine* was considerably higher than those selected from *Nature Medicine* caused by the discrepancies of published articles during the predefined review period. A further imbalance occurred regarding the in- and exclusion criteria for qualitative assessment. As *The New England Journal of Medicine* revealed a higher proportion of articles focusing on advanced statistical techniques not covered by the checklist, markedly more articles were excluded from this journal for the second part of the review. In this context, the development of additional review tools for specific, less frequently applied statistical methods would be desirable for future investigations. Due to insufficient information for the proper assessment of single checklist items in a considerable percentage of papers, the true error rate may be underestimated in our study.

Given the results of the present investigation, it should be recognized by all publishing authors and responsible journal editors, that [. . .] *the use of statistics in medical research may affect whether individuals live or die, whether their health is protected or jeopardized, and whether medical science advances or gets sidetracked. Therefore, all statistical practitioners, regardless of their training and occupation have a social obligation to perform their work in a professional and ethical manner.* (Ethical Guidelines for Statistical Practice, American Statistical Association, 1999)

[Received July 2006. Revised September 2006.]

REFERENCES

Altman, D. G. (1981), "Statistics and Ethics in Medical Research. Improving the Quality of Statistics in Medical Journals," *British Medical Journal*, 282, 44–47.

——— (1982), "Statistics in Medical Journals," *Statistics in Medicine*, 1, 59–71.

——— (1991), "Statistics in Medical Journals: Developments in the 1980s," *Statistics in Medicine*, 10, 1897–1913.

——— (1998), "Statistical Reviewing for Medical Journals," *Statistics in Medicine*, 17, 2661–2674.

——— (2000), "Statistics in Medical Journals: Some Recent Trends," *Statistics in Medicine*, 19, 3275–3289.

Altman, D. G., Gore, S. M., Gardner, M. J., and Pocock, S. J. (1983), "Statistical Guidelines for Contributors to Medical Journals," *British Medical Journal*, 286, 1489–1493.

Avram, M. G., Shanks, C. A., Dykes, M., Ronai, A. K., and Stiers, W. M. (1985), "Statistical Methods in Anesthesia Articles: An Evaluation of Two American Journals During two Six-Month Periods," *Anesthesia and Analgesia*, 64, 607–611.

Barron, J. P. (2006), "The Uniform Requirements for Manuscripts Submitted to Biomedical Journals Recommended by the International Committee of Medical Journal Editors," *Chest*, 129, 1098–1099.

Cardiel, M. H., and Goldsmith, C. H. (1995), "Type of Statistical Techniques in Rheumatology and Internal Medicine Journals," *Revista de Investigación Clinica*, 47, 197–201.

Clopper, C. J., and Pearson, E. S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404–413.

Colditz, G. A., and Emerson, J. D. (1985), "The Statistical Content of Published Medical Research: Some Implications for Biomedical Education," *Medical Education*, 19, 248–255.

Cooper, R. J., Schriger, D. L., and Close, R. J. H. (2002), "Graphical Literacy: The Quality of Graphs in a Large-Circulation Journal," *Annals of Emergency Medicine*, 40, 317–322.

Dar, R., Serlin, R., and Omer, H. (1994), "Misuse of Statistical Tests in Three Decades of Psychotherapy Research," *Journal of Consulting and Clinical Psychology*, 62, 75–82.

Davies, H. T. (1998), "Describing and Estimating: Use and Abuse of Standard Deviations and Standard Errors," *Hospital Medicine*, 59, 327–328.

Emerson, J. D., and Colditz, G. A. (1983), "Use of Statistical Analysis in the

New England Journal of Medicine," *New England Journal of Medicine*, 309, 709–713.

Felson, D. T., Cupples, L. A., and Meenan, R. F. (1984), "Misuse of Statistical Methods in Arthritis and Rheumatism. 1982 versus 1967-1968," *Arthritis and Rheumatism*, 27, 1018–1022.

García-Berthou, E., and Alcaraz, C. (2004), "Incongruence Between Test Statistics and *P* Values in Medical Papers," *BMC Medical Research Methodology*, 4, 13–17.

Gardenier, J. S., and Resnik, D. B. (2002), "The Misuse of Statistics: Concepts, Tools, and a Research Agenda," *Accountability in Research*, 9, 65–74.

Gardner, M. J., Altman, D. G., Jones, D. R., and Machin, D. (1983), "Is the Statistical Assessment of Papers Submitted to the *British Medical Journal* Effective?," *British Medical Journal*, 286, 1485–1488.

Gardner, M. J., Machin, D., and Campell, M. J. (1986), "Use of Check Lists in Assessing the Statistical Content of Medical Studies," *British Medical Journal*, 292, 810–812.

Gardner, M. J., and Bond, J. (1990), "An Exploratory Study of Statistical Assessment of Papers Published in the *British Medical Journal*," *Journal of the American Medical Association*, 263, 1355–1357.

Goodman, S. N., Altman, D. G., and George, S. L. (1998), "Statistical Reviewing Policies of Medical Journals," *Journal of General Internal Medicine*, 13, 753–756.

Goodman, N. W., and Hughes, A. O. (1992), "Statistical Awareness of Research Workers in British Anaesthesia," *British Journal of Anaesthesia*, 68, 321–324.

Gore, S. M., Jones, G., and Thompson, S. G. (1992), "The Lancet's Statistical Review Process: Areas for Improvement by Authors," *Lancet*, 340, 100–102.

Gore, S. M., Jones, I. G., and Rytter, E. C. (1977), "Misuse of Statistical Methods: Critical Assessment of Articles in *BMJ* from January to March 1976," *British Medical Journal*, 1, 85–87.

Horton, N. J., and Switzer, S. S. (2005), "Statistical Methods in the Journal," *New England Journal of Medicine*, 353, 1977–1979.

Huang, W., LaBerge, J. M., Lu, Y., and Glidden, D. V. (2002), "Research Publications in Vascular and Interventional Radiology: Research Topics, Study Designs, and Statistical Methods," *Journal of Vascular and Interventional Radiology*, 13, 247–255.

Kanter, M. H., and Taylor, J. R. (1994), "Accuracy of Statistical Methods in Transfusion: A Review of Articles from July/August 1992 through June 1993," *Transfusion*, 34, 697–701.

Kuzon, W. M., Urbanchek, M. G., and McCabe, S. (1996), "The Seven Deadly Sins of Statistical Analysis," *Annals of Plastic Surgery*, 37, 265–272.

MacArthur, R. D., and Jackson, G. G. (1984), "An Evaluation of the Use of Statistical Methodology in the*Journal of Infectious Diseases*," *Journal of Infectious Diseases*, 149, 349–354.

Marshall, S. W. (2004), "Testing with Confidence: The Use (and Misuse) of Confidence Intervals in Biomedical Research," *Journal of Science and Medicine in Sports*, 7, 135–137.

McCance, I. (1995), "Assessment of Statistical Procedures used in Papers in the *Australian Veterinary Journal*," *Australian Veterinary Journal*, 72, 322–328.

McGuigan, S. M. (1995), "The Use of Statistics in the *British Journal of Psychiatry*," *British Journal of Psychiatry*, 167, 683–688.

McKinney, W. P., Young, M. J., Hartz, A., and Bi-Fong Lee, M. (1989), "The Inexact use of Fisher's Exact Test in Six Major Medical Journals," *Journal of the American Medical Association*, 261, 3430–3433.

Menegazzi, J., Yealy, D., and Harris, J. (1991), "Methods of Data Analysis in the Emergency Medicine Literature," *American Journal of Emergency Medicine*, 9, 225–227.

Moher, D., Schulz, K. F., and Altman, D. G. (2001), "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials," *Lancet*, 357, 1191–1194.

Murray, G. D. (1991), "Statistical Guidelines for the *British Journal of Surgery*," *British Journal of Surgery*, 78, 782–784.

Olsen, C. H. (2003), "Review of the Use of Statistics in *Infection and Immunity*," *Infection and Immunity*, 71, 6689–6692.

Pocock, S. J., Hughes, M. D., and Lee, R. J. (1987), "Statistical Problems in the Reporting of Clinical Trials—A Survey of Three Medical Journals," *New England Journal of Medicine*, 317, 426–432.

Porter, A. M. (1999), "Misuse of Correlation and Regression in Three Medical Journals," *Journal of the Royal Society of Medicine*, 92, 123–128.

Reed III, J. F., Salen, P., and Bagher, P. (2003), "Methodological and Statistical Techniques: What do Residents Really Need to Know About Statistics?," *Journal of Medical Systems*, 27, 233–238.

Schor, S., and Karten, I. (1966), "Statistical Evaluation of Medical Manuscripts," *Journal of the American Medical Association*, 195, 1123–1128.

Thorn, M. D., Pulliam, C. C., Symons, M. J., and Eckel, F. M. (1985), "Statistical and Research Quality of the Medical and Pharmacy Literature," *American Journal of Hospital Pharmacy*, 42, 1077–1082.

Welch II, G. E., and Gabbe, S. G. (1996), "Review of Statistics Usage in the *American Journal of Obstetrics and Gynecology*," *American Journal of Obstetrics and Gynecology*, 175, 1138–1141.

White, S. J. (1979), "Statistical Errors in Papers in the *British Journal of Psychiatry*," *British Journal of Psychiatry*, 135, 336–342.

APPENDIX: CHECKLIST FOR STATISTICAL EVALUATION OF MEDICAL MANUSCRIPTS (1)

	Assessment		
	Error committed	*Unable to assess/ not clear*	*Application correct*
DESIGN OF STUDY			
1. Errors and Deficiencies related to randomization/blinding and selection of control groups			
Failure to use/report randomization (e.g. in a controlled trial/experiment)	O	O	O
Method of randomization/allocation to intervention not clearly stated (e.g. table of random numbers used)	O	O	O
Failure to report initial equality of baseline characteristics/comparability of study groups	O	O	O
Use of an inappropriate control group (heterogeneous, clearly not comparable material)	O	O	O
2. Errors and Deficiencies related to the design of the study			
Failure to report number of participants/observations (sample size)	O	O	O
Failure to report possible withdrawals from the study	O	O	O
No a priori sample size calculation/neglect of effect-size estimation; Power calculation	O	O	O
Inappropriate testing for equality of baseline characteristics (e.g., for initial statistical equality of groups)	O	O	O
DATA ANALYSIS			
3. Use of a wrong or suboptimal statistical test			
Incompatibility of statistical test with type of data examined	O	O	O
Unpaired tests for paired data (e.g., repeated observations analyzed as independent data) or vice versa	O	O	O
Inappropriate use of parametric methods (e.g., for data that are obviously non-Normal or skewed)	O	O	O
Use of an inappropriate test for the hypothesis under investigation	O	O	O
4. Multiple testing/multiple comparisons (Type I error inflation)			
Failure to include a multiple-comparison correction	O	O	O
Inappropriate post-hoc subgroup analysis (“shopping for statistical significant differences”)	O	O	O
5. Special errors with Student’s *t* test			
Failure to test and report that test assumptions were proven and met	O	O	O
Unequal sample sizes for paired *t* test	O	O	O
Improper multiple pair wise comparisons (without adjustment of alpha-level) of more than 2 groups	O	O	O
Use of an unpaired *t* test for paired data or vice versa	O	O	O
6. Special errors with χ^2-tests			
No Yates-continuity correction reported if small numbers	O	O	O
Use of chi-square when expected numbers in a cell are < 5	O	O	O
No explicit statement of the statistical null-hypothesis tested	O	O	O
p values obviously wrong	O	O	O
DOCUMENTATION			
7. Improper description of statistical tests			
Failure to specify/define all applied tests clearly and correctly	O	O	O
Wrong names for statistical tests	O	O	O
Referring to unusual/obscure methods without explanation or reference	O	O	O
Failure to specify which test was performed on a given set of data when more than one test was done	O	O	O
“Where appropriate” statement	O	O	O
8. Failure to define details of a test performed			
Failure to state number of tails	O	O	O
Failure to state if test was paired or unpaired	O	O	O
Failure to state in advance which values of *p* indicate statistical significance	O	O	O

continued on next page

	Assessment		
	Error committed	*Unable to assess/ not clear*	*Application correct*
PRESENTATION			
9. Inadequate (graphical or numerical) description/presentation of basic data (location, dispersion)			
Mean but no indication of variability of the data (failure to describe variability)	O	O	O
Giving Standard Error (SE) instead of Standard Deviation (SD) to describe/summarize study data	O	O	O
Failure to define $+/-$ notion for describing variability of the sample; unlabeled error bars	O	O	O
Use of arithmetic mean and SD to describe nonnormal or ordinal data	O	O	O
SE on undefined (or too small) sample sizes	O	O	O
10. Inappropriate/poor reporting of results			
Results given only as *p* values, no confidence intervals given for main effect size measures	O	O	O
CI given for each group rather than for the contrast	O	O	O
Numerical results and *p* values given to too many (or too few) decimal places (e.g. $p < 0.000000$)	O	O	O
"$p =$ NS," "$p < 0.05$," "$p > 0.05$" (or other arbitrary thresholds) instead of reporting exact *p* values	O	O	O
INTERPRETATION			
11. Wrong interpretation of results			
"nonsignificant" treated/interpreted as "no effect"/"no difference"	O	O	O
Marginal statistical significance (e.g., $p = 0.1$) treated as genuine effect	O	O	O
Drawing conclusions not supported by the study data	O	O	O
Significance claimed (or *p* values stated) without data analysis or statistical test mentioned	O	O	O
12. Poor interpretation of results			
Failure to consider CI's when interpreting "NS" differences (especially in small studies)	O	O	O
Disregard for Type II error when reporting nonsignificant results	O	O	O
Missing discussion of the problem of multiple significance testing if occurred	O	O	O
NOTES & COMMENTS			