AUSTRIAN JOURNAL OF STATISTICS
Volume 36 (2007), Number 2, 141–152

The Use of Statistics in Medical Research: A Comparison of Wiener Klinische Wochenschrift and Wiener Medizinische Wochenschrift

Alexander M. Strasak[1], Qamruz Zaman[1], Gerhard Marinell[2], Karl P. Pfeiffer[1], and Hanno Ulmer[1]

[1]Dept. of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Austria

[2]Inst. of Statistics, University of Innsbruck, Austria

Abstract: To evaluate the quantity and quality of the use of statistics in Austrian medical journals, all "original research" papers in No. 116/1-12 of *Wiener Klinische Wochenschrift* (WKW) and 153/1-24, 154/1-24 of *Wiener Medizinische Wochenschrift*(WMW) were screened for their statistical content. Types, frequencies and complexity of statistical methods applied were systematically recorded. A 46-item checklist was used to evaluate statistical quality for a subgroup of papers. 74.3% of WKW papers contained inferential methods beyond descriptive statistics. Only 43.7% of WMW papers employed methods of inferential statistics. There was a statistical significant difference regarding the use of statistical methods between the two journals ($p = 0.009$). In addition, complexity and sophistication of statistical analysis was considerable higher for WKW papers ($p = 0.02$). Statistical errors and deficiencies were identified in a large proportion of papers. Although inferential statistics were frequently identified in papers from WKW, only a minority of WMW research had analytical character. Types and frequencies of statistical errors identified, did not vary meaningful from findings of similar studies for a wide range of medical journals. There is reason to assume, that the journal impact-factor does not seem to be a powerful predictor for the statistical quality of published research.

Zusammenfassung: Zur Evaluierung von Quantität und Qualität des statistischen Inhalts in österreichischen medizinischen Fachzeitschriften wurden alle "Original Research" Publikationen in den No. 116/1-12 der Wiener Klinischen Wochenschrift (WKW) und 153/1-24, 154/1-24 der Wiener Medizinischen Wochenschrift (WMW) auf deren statistischen Inhalt analysiert. Statistische Methoden, deren Anwendungshäufigkeiten sowie Komplexitätsgrade wurden systematisch aufgezeichnet. Eine 46-Punkte-Checklist wurde zur Evaluierung der statistischen Qualität, für eine Untergruppe von Publikationen verwendet. 74.3% der WKW Publikationen enthielten statistische Methoden, die über deskriptive Statistik hinausgingen. Nur 43.7% der WMW Artikeln bedienten sich inferenzstatistischer Verfahren. Es konnte ein statistisch signifikanter Unterschied in der Verwendung statistischer Verfahren zwischen den beiden Zeitschriften festgestellt werden ($p = 0.009$). Darüber hinaus, war die Komplexität der statistischen Datenanalyse für WKW Publikationen signifikant höher als für WMW Publikationen ($p = 0.02$). Statistische Fehler und Defizite wurden in einem großen Anteil von Publikationen aus beiden Zeitschriften identifiziert. Obwohl inferenzstatistische Verfahren mit großer Häufigkeit in Publikationen der WKW festgestellt wurden, hatte nur eine Minderheit der WMW Papers analytischen Charakter. Da sich die gefundenen statistischen Fehler in Art und Häufigkeit nicht bedeutsam von den Ergebnissen ähnlicher Studien für eine große Zahl internationaler medizinischer Fachzeitschriften unterscheiden, darf angenommen werden, dass der Journal-Impact-Faktor zur Prognose der statistischen Qualität medizinischer Publikationen nur sehr bedingt geeignet ist.

Keywords: Statistics in Medicine, Techniques, Complexity, Errors.

1 Introduction

Statistical methods play vital roles in the scientific research process. Over the past decades, a great increase in the use of statistics has been documented, for a wide range of medical journals (Altman, 1982, 1991, 2000). Although, favored by the availability of manifold statistical software packages, a trend towards usage of more sophisticated techniques can be approved, there is also strong evidence, that in particular simple methods as t-tests or χ^2-tests remain in common application (Colditz and Emerson, 1985; Emerson and Colditz, 1983; Menegazzi et al., 1991; Cardiel and Goldsmith, 1995; Huang et al., 2002; Reed III et al., 2003).

The use of statistics in medical journals has been subjected to considerable review over the past four decades. There is wide compliance that standards are in general low, as a high proportion of published medical research contains statistical errors and deficiencies (Schor and Karten, 1966; Gore et al., 1976; Hoffmann, 1984; 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; Porter, 1999; Cooper et al., 2002; García-Berthou and Alcaraz, 2004). It seems safe to conclude that the problem is a serious one, as the inappropriate use of statistical analysis may lead to wrong conclusions or may weaken published research results. 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; Sheehan, 1980).

As a result, there was respectable effort from many medical journals, to enhance quality of statistics by adopting statistical guidelines for authors or by sharpening the statistical reviewing of incoming manuscripts (Gardner et al., 1983, Goodman et al., 1998; Gore et al., 1992; Altman, 1998; Altman et al., 1983; Murray, 1991). However, there is not much support for the idea that standards have largely improved over time, as also recent studies, although in general focussed to specific statistical affairs, point toward major problems (Cooper et al., 2002; García-Berthou and Alcaraz, 2004; Olsen, 2003; Marshall, 2004; Davies, 1998, Nagele, 2001; Freedman et al., 2001; Bezeau and Graves, 2001).

In this study we report on current usage of statistics in medicine, by reviewing original research papers from *Wiener Klinische Wochenschrift* (*The Middle European Journal of*

Medicine) and *Wiener Medizinische Wochenschrift*. The aim of the study was twofold: The first was to investigate the types and frequencies of statistical techniques applied, as well as the complexity of statistical analysis employed; 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, there is no comprehensive study, reviewing application of statistics for the two medical journals under scrutiny. Questions regarding their recent use therefore remain largely unanswered. The results of the study allow for an ongoing monitoring of possible trends in statistics usage and outline the most frequent errors and abuses, observed in a detailed quality assessment.

2 Material and Methods

All consecutive "*original research articles*" published during the first half of year 2004 in No. 116/1–12 of *Wiener Klinische Wochenschrift* and during years 2003 and 2004 in No. 153/1–24, No. 154/1–24 of *Wiener Medizinische Wochenschrift* were included for a Bibliometric analysis. Editorials, letters, case reports and review articles were excluded. Due to the small number of "original research" papers in *Wiener Medizinische Wochenschrift*, study period had to be extended adequately for this journal. There was a total of 35 papers in *Wiener Klinische Wochenschrift* (WKW) and 16 papers in *Wiener Medizinische Wochenschrift* (WMW).

All 51 papers were manually reviewed for their statistical content. Types and frequencies of statistical methods applied were systematically recorded and classified into 17 categories, similarly used by Emerson and Colditz (1983). Papers containing statistical analysis beyond descriptive statistics were further classified into "Basic Analysis" or "Advanced Analysis" according to complexity and sophistication of statistical analysis employed. For each paper, numbers of different statistical techniques were recorded.

A subgroup of 22 papers (WKW = 15, WMW = 7) was further selected for a detailed quality assessment of statistical methods employed. Assortment of papers for qualitative evaluation was done according to predefined inclusion criteria, with insistence on the use of inferential statistics and the use of at least one elementary statistical test in a paper. After detailed and critical examination of all sections, tables and figures, a standardized 46-item checklist was completed for each of the 22 papers, by the first author (A.M.S.). If an assessment was not clear or vague, the paper was independently reviewed by a second statistician (H.U.) and then assessed together. The 46-item checklist used for evaluation included multifaceted questions on statistical aspects of study design, statistical analysis, documentation of statistical methods applied, presentation and interpretation of study findings.

For the journals under investigation, two pre-specified hypotheses, regarding potential differences in the proportions of papers using (1) inferential methods and (2) advanced analysis, were tested. Statistical analysis was conducted by 2-tailed tests for linear trends in proportions and frequencies (Armitage, 1955) with a level of significance set at 0.05. Where useful, confidence intervals were computed by the method of Clopper and Pearson (1934).

3 Results

Table 1 shows the types and frequencies of statistical methods in all original research papers of No. 116/1-12 from WKW and No. 153/1-24, No. 154/1-24 from WMW. Of 35 papers analyzed from WKW, 74.3% (95%CI $56.7 - 87.5$) contained methods of inferential statistics. The corresponding number for WMW papers adds up to only 43.7%. There was a statistical significant difference regarding the use of statistical methods (No statistical methods/Descriptive statistics only/Inferential methods) between the journals ($p = 0.009$). Most frequently used inferential statistics in WKW were simple χ^2- and Fishers exact tests, identified in 12 of 35 papers (34.3%), closely followed by t-tests and non-parametric methods (e.g., U-tests, H-tests, Wilcoxon-tests), with a frequency of 28.6%, each. For WMW papers most commonly reported inferential methods were largely the same. Usage of confidence intervals only was identified in a considerable low proportion of papers reviewed.

Complexity and sophistication of statistical analysis, although in general quite moderate for both journals, was slightly more advanced for papers from WKW. Nevertheless, 34.3% of these papers also had to be classified "Basic Analysis", for constricting statistical evaluation to exclusively elementary techniques like t-tests, χ^2-tests, Fishers Exact tests, simple non-parametric tests, one-way ANOVA, or linear regression and correlation. 14 WKW papers (40.0%) reported usage of at least one more sophisticated method, beyond those listed above and therefore were classified "Advanced Analysis". The corresponding number for WMW papers equals to only 12.5%. There was a statistical significant difference regarding complexity of statistical analysis between the two journals when classifying papers into either (1) No/descriptive/unidentified methods, (2) Basic Analysis, or (3) Advanced Analysis ($p = 0.02$).

Table 2 and 3 show the types and frequencies of statistical errors and deficiencies, identified in a subsequent quality assessment of statistical methods employed. Most common error related to the design of a study was a failure to consider statistical sample size estimation or power calculation, especially in prospectively designed studies. Three of 15 papers from WKW (20.0%) contained usage of wrong statistical tests, either because of incompatibility of test with data examined, inappropriate use of parametric methods, or use of an inappropriate statistical test for the scientific hypothesis under investigation. The correspondent proportion for WMW papers, although in general quite moderate sophistication of statistical analysis, equals to 42.9%, and therefore was considerable higher. Because of intense and persistent deficiencies in documentation of statistical methods employed, it was in general fairly difficult to determine accuracy and appropriateness of statistical analysis. There was a high rate of papers with checklist-assessment "unable to assess/not clear" (data not shown).

Other frequently observed statistical abuses were usage of undefined +/– notions or unlabeled error bars for describing variability of data and inaccurate reporting of arbitrary thresholds, instead of specifying exactly obtained p-values. Common statistical errors related to interpretation of study findings were the erroneous discussion of non-significant results as "no effect/no difference" and neglect of multiple testing problems, commonly associated with multiple study endpoints.

Table 1: Types, frequencies and complexity of statistics

	WKW		WMW		
	($n = 35$)		($n = 16$)		p Value[a]
Types and frequencies of statistical methods[b]	n	%	n	%	
No statistical methods	1	2.9	4	25.0	
Descriptive statistics only	8	22.9	5	31.3	
Inferential methods	26	74.3	7	43.7	0.009
t-tests	10	28.6	2	12.5	
Contingency table analysis					
Basic (χ^2-, Fishers exact test)	12	34.3	5	31.3	
Advanced	1	2.9	0	0.0	
Non-parametric tests	10	28.6	4	25.0	
Analysis of variance					
Basic (one-way ANOVA)	2	5.7	0	0.0	
Advanced	1	2.9	0	0.0	
Correlation coefficients	8	22.9	2	12.5	
Regression					
Basic (simple-linear regression)	2	5.7	0	0.0	
Advanced	7	20.0	1	6.3	
Epidemiologic methods	5	14.3	1	6.3	
Survival analysis	2	5.7	0	0.0	
Other methods	2	5.7	0	0.0	
Confidence intervals	5	14.3	2	12.5	
Complexity of statistical analysis					
No. of different inferential methods					
Only 1 method	5	14.3	1	6.3	
2 or 3 methods	13	37.1	3	18.8	
4 or 5 methods	6	17.1	3	18.8	
More than 5 methods	2	5.7	0	0.0	
No/descriptive/unidentified methods	9	25.7	9	56.3	
Basic analysis[c]	12	34.3	5	31.3	
Advanced analysis[d]	14	40.0	2	12.5	0.02

[a] determined by 2-tailed tests for linear trends in proportions and frequencies (Armitage, 1955).

[b] 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%. A full explanation for the categories listed is given by Emerson and Colditz (1983).

[c] t-tests, contingency table analysis basic, non-parametric tests, ANOVA basic, correlation coefficients, regression Basic.

[d] 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 analysis" was obligatory.

4 Discussion

The implications of the study at hand are twofold: First the results give up to date evidence for the widespread use of statistics, also in the *Middle European Journal of Medicine*. As

Table 2: Statistical errors, flaws and deficiencies related to the design of a study and statistical analysis[a]

	WKW ($n = 15$)		WMW ($n = 7$)	
Category	n	%	n	%
Design of study				
No sample size calculation/power calculation (overall)	11	73.3	4	57.1
Prospective study design	4	26.7	2	28.6
Retrospective study design	4	26.7	2	28.6
Study design not classifiable[b]	3	20.0	0	0.0
Data analysis				
Use of a wrong statistical test	3	20.0	3	42.9
Incompatibility of statistical test with type of data examined	2	13.3	1	14.3
Inappropriate use of parametric methods	1	6.7	1	14.3
Unpaired tests for paired data or vice versa	0	0.0	1	14.3
Failure to include a multiple-comparison correction/α-level correction	3	20.0	1	14.3
Special errors with Student's t-test				
Failure to proof/report that test assumptions are not violated	5	33.3	2	28.6[c]
Unequal sample sizes for paired t-test	1	6.7	0	0.0
Special errors with χ^2-tests				
No Yates correction if small numbers	2	13.3	0	0.0
Use of chi-square when expected numbers in a cell are < 5	1	6.7	2	28.6
p-values obviously wrong[d]	0	0.0	1	14.3

[a] papers with checklist assessment "unable to assess/not clear" not shown.

[b] papers did not contain sufficient information to clearly classify design of study.

[c] as only 2 papers reported usage of t-tests, the actual proportion of papers, neglecting t-test assumptions, equals to 100%.

[d] recalculation of p-values, as only category, was not done systematically and for all p-values presented, but only in cases, when clear discrepancies between data and test results could be identified.

nearly 75.0% of papers reviewed had analytical character, using some kind of inferential methods, the results of the present study for papers in the *Middle European Journal of Medicine* correspond widely to findings of earlier studies for a wide range of medical journals, attesting similar proportions (Menegazzi et al., 1991; Cardiel and Goldsmith, 1995; Huang et al., 2002; Reed III et al., 2003). This does not necessarily hold for papers from WMW, as 9 of 16 research papers reviewed, were purely descriptive, without any analytical power. Thus, it eventually should be reconsidered by the editors, if their possible impact on medical research justifies their frequency.

As a second implication of the study, it can be concluded that statistical errors and deficiencies seem to be common also in the *Middle European Journal of Medicine* and *Wiener Medizinische Wochenschrift*. 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 a closer editorial attention to statistical methodology, starting at the pre-publication phase of a manuscript, should emphatically be considered by the editors to raise standards and thereby, possibly improve journal impact-factors.

Contrariwise, it can be argued with caution, that the journal-impact-factor does not

Table 3: Statistical errors, flaws and deficiencies related to documentation, presentation and interpretation[a]

	WKW ($n = 15$)		WMW ($n = 7$)	
Category	n	%	n	%
Documentation				
Failure to specify/define all statistical tests used clear and correctly	12	80.0	6	85.7
Failure to state number of tails	12	80.0	6	85.7
Failure to state if test was paired or unpaired	5	33.3	0	0.0
Failure to specify which test was performed on a given set of data[b]	4	26.7	1	14.3
Wrong names for statistical tests	2	13.3	2	28.6
Failure to state which values of p indicate statistical significance	8	53.3	3	42.9
Presentation				
"Mean" but no indication of variability of data[c]	1	6.7	0	0.0
Giving standard error (se) instead of sd for statistical description	1	6.7	0	0.0
Failure to define +/– notion for describing variability; use of unlabelled error bars	3	20.0	2	28.6
No confidence intervals for main effect size measures presented	10	66.7	5	71.4
$p =$ NS, $p < 0.05$, $p > 0.05$ etc. instead of reporting exact p-values	7	46.7	5	71.4
Interpretation[d]				
"non significant" treated/interpreted as "no effect"/"no difference"	3	20.0	1	14.3
Significance claimed without data analysis or statistical test mentioned	1	6.7	0	0.0
Disregard for type II error when reporting non-significant results	2	13.3	0	0.0
No discussion of problem of multiple significance testing if occurred	4	26.7	1	14.3

[a] papers with checklist assessment "unable to assess/not clear" not shown.

[b] statistical assessment for this category only was possible when more than one type of statistical test was reported and performed.

[c] "mean" corresponds to any kind of statistical measure of central tendency.

[d] 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.

seem to be a meaningful predictor for statistical quality of published medical research, as types and frequencies of statistical errors and deficiencies identified, although generally concerning, did not differ substantially from earlier results, found in similar studies for other, partially high-impact medical journals (Schor and Karten, 1966; Gore et al., 1976; 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; Olsen, 2003; Marshall, 2004; Davies, 1998; Nagele, 2001).

Moreover, it should be acknowledged, 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 a considerable proportion of papers from both journals, as many authors failed to specify accurately all statistical tests used for generating the p-values presented. As also stressed in an early study from Pocock et al. (1987) for three major medical journals, more emphasis should be given to the magnitude of treatment differences and to statistical estimation techniques as confidence intervals, than to solely rely on uncritical significance testing.

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.

Altman, D. G. (1982). Statistics in medical journals. *Statistics in Medicine*, *1*, 59-71.

Altman, D. G. (1991). Statistics in medical journals: Developments in the 1980s. *Statistics in Medicine*, *10*, 1897-1913.

Altman, D. G. (1998). Statistical reviewing for medical journals. *Statistics in Medicine*, *17*, 2661-2674.

Altman, D. G. (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.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375-386.

Bezeau, S., and Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, *23*, 399-406.

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.

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.

Freedman, K. B., Back, S., and Bernstein, J. (2001). Sample size and statistical power of randomised, controlled trials in orthopaedics. *Journal of Bone and Joint Surgery*, *83*, 397-402.

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., 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.

Gore, S. M., Jones, I. G., and Rytter, E. C. (1976). Misuse of statistical methods: critical assessment of articles in bmj from january to march 1976. *British Medical Journal*, *1*, 85-87.

Gore, S. M., Jones, I. G., and Thompson, S. G. (1992). The Lancet's statistical review process: areas for improvement by authors. *Lancet*, *340*, 100-102.

Hoffmann, O. (1984). Application of statistics and frequency of statistical errors in articles in Acta Neurochirurgica. *Acta Neurochirurgica*, *71*, 307-315.

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.

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.

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.

Murray, G. D. (1991). Statistical guidelines for the British Journal of Surgery. *British Journal of Surgery*, *78*, 782-784.

Nagele, P. (2001). Misuse of standard error of the mean (sem) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia*, *90*, 514-516.

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.

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.

Sheehan, T. J. (1980). The medical literature – Let the reader beware. *Archives of Internal Medicine*, *140*, 472-474.

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.

Corresponding Authors' address:
Alexander M. Strasak
Department of Medical Statistics, Informatics and Health Economics
Innsbruck Medical University
Schoepfstrasse 41
A-6020 Innsbruck, Austria
Tel. +43 (512) 507 3221
Fax +43 (512) 507 2711
E-mail: alexander.strasak@i-med.ac.at

Appendix. Checklist for Statistical Evaluation of Medical Manuscripts

Statistical Checklist (1)

	Assessment		
	A	B	C
Design of Study			
1 Errors & 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 & 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 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

A = error committed, B = unable to assess/not clear, C = application correct

Statistical Checklist (2)

	Assessment		
	A	B	C
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
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 non-normal 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			
“non significant” 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 non-significant results	o	o	o
Missing discussion of the problem of multiple significance testing if occurred	o	o	o

Notes & Comments

A = error committed, B = unable to assess/not clear, C = application correct