International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 16(4): 202–207 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.225

Inconsistencies between reported test statistics and *p*-values in two psychiatry journals

DAVID BERLE,[1] VLADAN STARCEVIC[2]

1 Nepean Anxiety Disorders Clinic, Sydney West Area Health Service, Penrith, NSW, Australia
2 University of Sydney and Nepean Hospital, Sydney/Penrith, NSW, Australia

Abstract

A recent survey of the British Medical Journal (BMJ) and Nature revealed that inconsistencies in reported statistics were common. We sought to replicate that survey in the psychiatry literature. We checked the consistency of reported t-test, F-test and χ^2-test values with their corresponding p-values in the 2005 issues of the Australian and New Zealand Journal of Psychiatry (ANZJP) and compared this with the issues of the ANZJP from 2000, and with a similar journal, Acta Psychiatrica Scandinavica (APS). A reported p-value was 'inconsistent' if it differed (at its reported number of decimal places) from our calculated p-values (using three different software packages), which we based on the reported test statistic and degrees of freedom. Of the 546 results that we checked, 78 (14.3%) of the p-values were inconsistent with the corresponding degrees of freedom and test statistic. Similar rates of inconsistency were found in APS and ANZJP, and when comparing the ANZJP between 2000 and 2005. The percentages of articles with at least one inconsistency were 8.5% for ANZJP 2005, 9.9% for ANZJP 2000 and 12.1% for APS. We conclude that inconsistencies in p-values are common and may reflect errors of analysis and rounding, typographic errors or typesetting errors. Suggestions for reducing the occurrence of such inconsistencies are provided. Copyright © 2008 John Wiley & Sons, Ltd.

Key words: statistics, results, conclusions, errors, inconsistencies

Introduction

There is a substantial literature documenting the inappropriate or erroneous use of statistical procedures in medical research (Avram et al., 1985; Cruess, 1989; Emerson and Colditz, 1983; Felson et al., 1984; Gore et al., 1977; MacArthur and Jackson, 1984; Olsen, 2003; Vrbos et al., 1993). Estimates of the occurrence of errors vary, but the most comprehensive reviews suggest that around 50% of articles contain at least minor statistical errors (Felson et al, 1984; Gore et al., 1977; Olsen, 2003; Vrbos et al., 1993).

In the psychiatry literature, McGuigan (1995) found that 40% of the 164 papers he surveyed from the *British Journal of Psychiatry* contained at least minor statistical errors. In reviewing papers from the *American Journal of Psychiatry* and the *Australian and New Zealand Journal of Psychiatry* (ANZJP), Davies (1987) found a total of 76 'errors' of design and analysis in his sample of 29 papers, 37 (49%) of which could be considered statistical errors.

However, quantifying rates of statistical errors is not easy. Authors have applied differing levels of stringency in defining what they consider to be an error, and there remains debate among statisticians as to what constitutes an error (Hand and Sham, 1995). For instance, failure to adjust the significance level for multiple comparisons is apparently a frequently occurring 'error' (Davies, 1987; McGuigan, 1995) despite a lack of consensus among experts as to when adjustments should be made and to what extent (Hall and Bird, 1985; Murphy, 2004; Perneger, 1998). Moreover, the consequences of statistical errors are often not sufficient to lead to altered conclusions. Nonetheless, reviews of mistakes and inconsistencies are likely to

underestimate the true occurrence of errors, given that many errors cannot be detected from perusal of the papers themselves and that most reviews of statistical reporting have surveyed prestigious journals where rigorous peer review may be more likely (Hand, 1985).

Although debate and a lack of consensus surrounds the appropriate choice and application of statistical tests, simple calculations can be performed to establish rates of consistency between reported test statistics, degrees of freedom and *p*-values (García-Berthou and Alcaraz, 2004). However, *p*-values are not a favored way of reporting results, as the magnitude of an effect is not suggested by a *p*-value (Cohen, 1994). Nevertheless, in the absence of effect sizes or confidence intervals, an estimate of the effect size can often be calculated from a correctly reported test statistic and/or *p*-value. Furthermore, *p*-values remain the most popular way of reporting results in the psychiatric literature, and many medical journals (e.g. *Japanese Journal of Clinical Oncology* and *The Canadian Journal of Psychiatry*) require authors to report the test statistic, degrees of freedom and exact *p*-value. Finally, the accuracy of *p*-values can be easily verified by readers whereas checking the accuracy of confidence intervals is more difficult. García-Berthou and Alcaraz reported that '11.6% (21 of 181) and 11.1% (7 of 63) of the statistical results published in *Nature* and BMJ [*British Medical Journal*] respectively during 2001 were incongruent, probably mostly due to rounding, transcription, or type-setting errors. At least one such error appeared in 38% and 25% of the papers in *Nature* and BMJ respectively' (García-Berthou and Alcaraz, 2004).

To the best of our knowledge, there have been no studies in the past decade that investigated inconsistencies between *p*-values and test statistics in the psychiatry literature. Thus, we sought to determine how frequently inconsistencies occurred in papers published in the ANZJP and in the *Acta Psychiatrica Scandinavica* (APS). By applying the approach of García-Berthou and Alcaraz (2004), we checked the consistency of *p*-values with common statistical tests reported in recent issues of the ANZJP, and compared this with the consistency of *p*-values in articles published in the same journal in 2000, and with the rates of inconsistency in APS. A reported *p*-value was considered to be 'inconsistent' if it differed (at its reported number of decimal places) from our calculated *p*-values, which we based on the reported test statistic and degrees of freedom. We chose to survey the ANZJP and APS as these are general psychiatry journals that publish papers on various topics, with a similar number of articles per year. Also, we compared ANZJP in 2005 and 2000 to ascertain whether there were any changes over time in the frequency with which inconsistencies in reported *p*-values occurred.

Method

We searched volume 39 of the ANZJP (2005; 12 issues, excluding supplements 1 and 2), volume 34 of the ANZJP (2000; six issues, excluding supplements 1 and 2), and volumes 111 and 112 of APS (2005; 12 issues, excluding supplements s426–s428) for statistical results that included the test statistic, degrees of freedom and *p*-value. To limit the magnitude of the task, we chose to focus on Chi-square (χ^2), Student's *t* and Analysis of Variance (ANOVA, MANOVA, and ANCOVA) tests, which are among the most commonly reported statistics (Altman and Bland, 1990; McGuigan, 1995). Other tests that are based on similar distributions, for instance, the χ^2 goodness-of-fit statistic in logistic regression, were not included in our survey. We searched the PDF files of original articles, review articles, and case series using Adobe Acrobat's 'search' function. We chose to search volume 34 (2000) of the ANZJP (in addition to volume 39) to allow an estimate of whether the rates of inconsistency have changed across time. Volume 34 was searched in preference to earlier volumes, as computerized searches of full-text articles were not possible for volumes of 10, 15 or 20 years ago.

If the test statistic and the degrees of freedom for a reported result are known, the *p*-value can be calculated using most statistical programs. In considering the consistency of *p*-values, we did not impose a predetermined set number of decimal places against which we checked all *p*-values. Instead, we checked *p*-values with the number of decimal places reported in each paper. We did not include instances where inexact *p*-values were reported (e.g. $p < 0.05$ or $p > 0.05$), as only gross inconsistencies can be detected for these.

Errors and inaccuracies in statistical software are not uncommon (Dallal, 1988; Knüsel, 1998) and can be a function of the algorithm used by the program to compute the statistic, rounding error within the software or computer hardware, or truncation error, which results in errors of approximation in the output (McCullough, 1998). So, we checked each reported statistic using three different software packages: SPSS

DOI: 10.1002/mpr

12.0.1 for Windows, Microsoft Excel 2000 and the freeware NCSS Probability Calculator for Windows. Where reported *p*-values were inconsistent with the computations of all three packages, we checked to ensure that the reported *p*-value was not a result of rounding in the test statistic of the original paper. For instance, the result '$\chi^2 = 0.4$, $df = 1$, $p = 0.52$' in volume 39(6), p. 476 of the ANZJP, is not necessarily inconsistent with the precise calculated value of $p = 0.5271$, considering that the rounded χ^2 value of 0.4 corresponds to the values from 0.35 to 0.44. There were occasional instances where, of the three statistical software packages, one produced a *p*-value which varied from the other two (at less than six decimal places). Such discrepancies were usually very small in magnitude (less than 0.001) and in the few instances where this occurred, we assumed that the value produced by the two consistent software packages was the correct value. In only one case did the reported value and the values calculated by the three software packages all differ from each other. In that case the reported *p*-value ($p = 0.49$ in volume 112(4) p. 268 of APS) was discrepant from all three of our calculated values (which, however, were each 0.43 when rounded to two decimal places) and thus, was considered to be an inconsistency.

Results

The 12 issues surveyed in the 2005 ANZJP included 118 articles, of which 32 (27.1%) reported χ^2, *t*, or *F* statistics with degrees of freedom and a *p*-value. Of the 260 χ^2, *t* and *F* statistics retrieved, 155 (59.6%) included a precise *p*-value, as opposed to an inequality (e.g. $p < 0.05$) or 'ns' (not significant). Of the six issues of the ANZJP in 2000 that were surveyed, 25 (22.5%) of 111 articles reported χ^2, *t*, or *F* statistics with degrees of freedom and a *p*-value. From these articles, 297 statistical results were retrieved, of which 173 (58.2%) included a precise *p*-value. In the 2005 APS, 39 (33.6%) of 116 articles reported χ^2, *t*, or *F* statistics with degrees of freedom and a *p*-value. From these articles, 445 statistical results were retrieved, of which 218 (49.0%) included a precise *p*-value. The mean number of χ^2, *t*, or *F*-test results reported per article (across all surveyed articles), was 2.9.

The numbers and percentages of inconsistent *p*-values are reported in Table 1. The 2000 and 2005 issues of the ANZJP had similar rates of inconsistency (13.9% and 14.8% respectively). Table 1 also shows that, in 2005, the ANZJP and the APS had similar rates of inconsistency (14.8% and 14.2% respectively). The greatest proportion of inconsistencies (18.0%) occurred

Table 1. Number and percentage of statistical inconsistencies by test, journal and year

		Number reported	Number inconsistent	Percent inconsistent
2000 ANZJP	*t* test	24	2	8.3%
	F test	70	5	7.1%
	χ^2 test	79	17	21.5%
	Total number of tests	173	24	13.9%
2005 ANZJP	*t* test	24	4	16.7%
	F test	70	14	20.0%
	χ^2 test	61	5	8.2%
	Total number of tests	155	23	14.8%
2005 APS	*t* test	57	3	5.3%
	F test	84	11	13.1%
	χ^2 test	77	17	22.1%
	Total number of tests	218	31	14.2%
Total	*t* test	105	9	8.6%
	F test	224	30	13.4%
	χ^2 test	217	39	18.0%
	Total number of tests	546	78	14.3%

ANZJP, *Australian and New Zealand Journal of Psychiatry.*
APS, *Acta Psychiatrica Scandinavica.*

DOI: 10.1002/mpr

with χ^2 tests. There was a slight tendency to report higher *p*-values: 60.8% of inconsistent *p*-values were higher than our calculated values.

The frequency and proportion of inconsistencies out of the total number of articles, and estimates of the number of potentially affected statistical decisions (using $p < 0.05$ as an arbitrary type I error rate), are contained in Table 2. Of all the papers surveyed, 10.1% contained at least one inconsistency and 2.6% would potentially have at least one erroneous statistical significance decision. Of the articles that contained at least one inconsistency, a high proportion contained two (20.6%), three (17.7%), or four or more (17.7%) inconsistencies.

Our methodology did not allow the types of errors leading to each inconsistency to be identified. However, many of the inconsistencies were probably due to typographic or typesetting errors. Examples include: '$p < 001$' (p. 806 of ANJZP, Vol. 34(5)), '$p < 05$' (p. 450 of ANZJP, Vol. 34(3)), '$c^2 = 8.2$' (p. 294 of ANZJP, Vol. 34(2)) and omission of a '0' in the *p*-value of: '$t = 3.4$, $df = 15$, $p = 0.04$' (p. 609 of ANZJP, Vol. 39(7); correct *p*-value = 0.004). Apparent rounding errors were also common, even after the rounding of the test statistic was considered: '$\chi^2 = 3.59$, $df = 1$, $p = 0.05$' (consistent *p*-value = 0.058; p. 789 of ANZJP, Vol. 34(5)) and '$F = 0.25$, $df = 3{,}427$, $p = 0.87$' (consistent *p*-value = 0.861; p. 82 of ANZJP, Vol. 34(1)).

Discussion

The present survey revealed that inconsistencies in *p*-values were common in the ANZJP and APS and that they did not vary greatly across time for ANZJP, or between these two journals. Many papers contained at least one inconsistency, and of those, more than half contained multiple inconsistencies. However, assuming a type I error rate of 0.05, in only 2.61% of papers would at least one statistical decision have been affected. Inconsistent *p*-values tended to be higher than our calculated values; however, we could not determine whether this was due to errors in rounding or to other factors, such as typographic errors.

Of the three statistical tests we surveyed, the highest proportion of inconsistencies occurred with χ^2 tests. Although the reason for this is unclear, this finding may indicate a need to be particularly attentive when using and reporting the results of χ^2 tests.

The inconsistency rate (14.3% overall) was slightly higher than that of the general medical and science literature surveyed by García-Berthou and Alcaraz (2004), who found that 11.1% of results in the *British Medical Journal* (BMJ) and 11.6% of results in *Nature*

Table 2. Proportions of surveyed papers where an inconsistent statistic was reported and proportions of papers where the statistical conclusion may have been affected

	Number of articles[1]	Number of articles with inconsistencies	Percent of articles with inconsistencies	Number of articles with at least one potentially affected statistical conclusion[2]	Percent of articles with at least one statistical conclusion affected
ANZJP 2000 (6 issues)	111	11	9.9%	3	2.7%
ANZJP 2005 (12 issues)	118	10	8.5%	5	4.2%
APS 2005 (12 issues)	116	14	12.1%	1	0.9%
Total	345	35	10.1%	9	2.6%

[1] This refers to the number of articles surveyed and does not include letters to the editor, book reviews, editorials or articles that were published in supplements. It does however, include case series and review articles.
[2] This refers to the number of times that an inconsistency potentially resulted in an erroneous statistical significance decision. It assumes that all authors used a significance criterion of $\alpha = 0.05$ and does not allow for multiple comparisons.
ANZJP, *Australian and New Zealand Journal of Psychiatry*.
APS, Acta Psychiatrica Scandinavica.

DOI: 10.1002/mpr

were inconsistent. This may be due to the fact that our survey focused only on three particular test distributions (*t*, *F* and χ^2), whereas García-Berthou and Alcaraz (2004) included all statistics where a statistic value, degrees of freedom and *p*-value were reported. Perhaps inconsistencies are more likely to occur with *t*, *F* and χ^2 tests. It is also possible that the BMJ and *Nature* have particularly rigorous reviewing and editorial processes that are more likely to detect discrepancies. Compared with the García-Berthou and Alcaraz (2004) sample, relatively few of our surveyed papers contained at least one inconsistency. Again, this may have been due to our decision to focus only on three particular test statistics, as well as to variations in the types of papers published in each journal. Our reduced rates of inconsistency per article is not necessarily a consequence of fewer tests per article in the ANZJP and APS. As already noted, of the articles that we surveyed, an average of 2.9 χ^2, *t*, or *F* tests were reported per article. Such multiple testing is troublesome if many of these articles contained correlated tests, redundant tests, or unnecessary tests. Finally, we might have found more inconsistencies had the number of statistical results reporting precise *p*-values been greater. Our finding that only slightly more than one half of all statistical reports included a precise *p*-value is in itself of some concern.

It is noteworthy that despite increased use of statistical software packages and increased attention by editors and reviewers to statistical reporting in recent years, we did not find appreciably different rates of inconsistency across the 5-year period from 2000 to 2005, at least in so far as ANZJP is concerned. This suggests that authors, editors and readers should avoid complacency in considering the accuracy of reported results.

A limitation of the present methodology is that even though it allows inconsistencies in reported statistical results to be accurately quantified, it is difficult to generalize these findings to all types of statistical result, and form conclusions about the reporting of statistics where imprecise *p*-values are given, and where degrees of freedom are not reported. An additional limitation is that our approach did not allow the type of error (be it a statistical calculation error, a rounding error, or a typesetting error) that led to each inconsistency to be determined. Finally, the consistency of statistics which describe the range or magnitude of effect, such as confidence intervals, could not be established using this approach.

Despite these limitations, the frequency of inconsistency found in the current survey is of concern. This is even more so, if one assumes that many readers of professional journals pay most attention to *p*-values, considering them to be the single most important result of statistical analyses and not having time, interest or expertise to look into other aspects of the statistical procedures. We surveyed the consistency of reporting of three common and relatively simple statistical procedures. This represents only one part of the analysis and reporting process. Errors may also occur at other stages in the course of preparing the manuscript: during data entry, in the selection of statistical tests, in the analysis of data or in the use of advanced statistical procedures. It is important for checks to be performed at each of these stages, especially if the paper passes between numerous contributors and undergoes many modifications during the preparation process. It is difficult or impossible for reviewers or readers to detect errors during various stages of research.

Authors, reviewers and editors can all play a part in reducing inconsistencies and errors in statistical reporting. Authors need to be cautious when rounding digits and need to check their statistics for typographic errors. Given occasional inconsistencies between software programs, replicating statistical results in a second software package is recommended. It is important for authors to report exact *p*-values to help reduce the perpetuation of '$p < 0.05$' as the only criterion for rejecting hypotheses (Wright, 2003). Authors should also adhere to established reporting guidelines and recommendations (e.g. American Psychological Association, 2001; Altman et al., 2001) by reporting confidence intervals (which also provide information about the precision of findings; Cumming and Finch, 2001) and effect sizes. Reviewers might be encouraged to apply some of the simple techniques we have described here to check reported *p*-values. This will be facilitated if authors report test statistics and *p*-values to the same number of decimal places, so that a reviewer can be reasonably confident that any inconsistency is not due to rounding of the test statistic. We believe that reporting to three decimal places will be appropriate in most circumstances, especially as it ensures that rounding is much less likely to affect the significance of a value (e.g. we suggest reporting $p = 0.049$ rather than $p = 0.05$). Finally, editors can promote quality reporting practices in their instructions to authors sections, such as the reporting of confidence intervals, or at the very least,

DOI: 10.1002/mpr

the reporting of exact *p*-values. To increase the accountability for each stage of analysis and reporting, editors might ask that authors specify which individuals were responsible for each stage of data collection, analysis and reporting (Balon, 2005), and this information might also appear in the final published article, as is now standard practice in the *Journal of the American Medical Association* (JAMA) and *The Lancet*.

References

Altman DG, Bland JM. Improving doctors' understanding of statistics. J R Stat Soc (Ser A) 1990; 154: 223–67.

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Int Med 2001; 134: 663–94.

American Psychological Association. Publication Manual. Washington, DC: American Psychological Association, 2001.

Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg 1985; 64: 607–11.

Balon R. By whom and how is the quality of research data collection assured and checked? Psychother Psychosom 2005; 74: 331–35.

Cohen J. The earth is round ($p < 0.05$). Am Psychol 1994; 49: 997–1003.

Cruess DF. Review of the use of statistics in the *American Journal of Tropical Medicine and Hygiene* for January–December 1988. Am J Trop Med Hyg 1989; 41: 619–26.

Cumming G, Finch S. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. Educ Psychol Meas 2001; 61: 532–74.

Dallal GE. Statistical microcomputing – like it is. Am Stat 1988; 42: 212–16.

Davies J. A critical survey of scientific methods in two psychiatry journals. Aust N Z J Psychiatry 1987; 21: 367–73.

Emerson JD, Colditz GA. Use of statistical analysis in the *New England Journal of Medicine*. N Engl J Med 1983; 309: 709–13.

Felson DT, Cupples LA, Meenan RF. Misuse of statistical methods in *Arthritis and Rheumatism*. 1982 versus 1967–68. Arthritis Rheum 1984; 27: 1018–22.

García-Berthou EC, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Med Res Methodol 2004; 4: 13.

Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. BMJ 1977; 1: 85–7.

Hall W, Bird KD. The problem of multiple inference in psychiatric research. Aust N Z J Psychiatry 1985; 19: 265–74.

Hand D, Sham P. Improving the quality of statistics in psychiatric research. Br J Psychiatry 1995; 167: 689–91.

Hand DJ. The role of statistics in psychiatry. Psychol Med 1985; 15: 471–6.

Knüsel L. On the accuracy of statistical distributions in Microsoft Excel 97. Comput Stat and Data Anal 1988; 26: 375–7.

MacArthur RD, Jackson GG. An evaluation of the use of statistical methodology in the *Journal of Infectious Diseases*. J Infect Dis 1984; 149: 349–54.

McCullough BD. Assessing the reliability of statistical software: Part I. Am Stat 1998; 52: 358–66.

McGuigan SM. The use of statistics in the *British Journal of Psychiatry*. Br J Psychiatry 1995; 167: 683–8.

Murphy JR. Statistical errors in immunologic research. J Allergy Clin Immunol 2004; 114: 1259–63.

Olsen CH. Review of the use of statistics in *Infection and Immunity*. Infect Immun 2003; 71: 6689–92.

Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998; 316: 1236–8.

Vrbos LA, Lorenz MA, Peabody EH, McGregor M. Clinical methodologies and incidence of appropriate statistical testing in orthopaedic spine literature. Are statistics misleading? Spine 1993; 18: 1021–9.

Wright DB. Making friends with your data: improving how statistics are conducted and reported. Br J Educ Psychol 2003; 73: 123–36.

Correspondence: David Berle, Nepean Anxiety Disorders Clinic, Department of Psychological Medicine, Nepean Hospital, PO Box 63, Penrith NSW 2751, Australia.
Telephone +61 2 4731 6504
Fax +61 2 4731 5279
Email: dberle@bigpond.net.au