International Psychogeriatrics (2004), 16:4, 389–396 © 2004 International Psychogeriatric Association
DOI: 10.1017/S1041610204000821 Printed in the United Kingdom

STATISTICAL AUDIT

Statistical audit of original research articles in *International Psychogeriatrics* for the year 2003

John T. Chibnall
Associate Professor of Psychiatry, Saint Louis University School of Medicine, St. Louis, MO, U.S.A.
Statistical Advisor to the Editor-in-Chief, International Psychogeriatrics

ABSTRACT

Background: At the request of the Editor of *International Psychogeriatrics*, a statistical audit of all papers published in the journal during 2003 was undertaken by the statistical advisor to *International Psychogeriatrics*.

Method: Only research papers using inferential statistical techniques were assessed and only the statistical elements of these papers were evaluated. The following issues were addressed: did the authors report a power calculation or address power issues? Did the authors report an appropriate effect size indicator? When multiple univariate statistical tests were used was a correction for type 1 error employed? Did authors demonstrate the adequacy of the data analyzed for the statistical tests employed? Were sufficient details reported to enable an evaluation of the statistical analyses and reported results?

Results: Twenty papers published during 2003 were suitable for analysis. None addressed power issues. About half reported an effect size indicator and about half adjusted the statistical analysis for the effects of multiple univariate statistical comparisons. Few demonstrated the adequacy of the data being analyzed and few provided sufficient detail to evaluate the statistical analyses and reported results. Most papers used the right statistic in the right way.

Conclusion: The statistical quality of articles published in *International Psychogeriatrics* could be improved by attention to a few relatively fundamental issues.

Key words: Statistical audit, *International Psychogeriatrics*, 2003

Correspondence should be addressed to: John. T. Chibnall, St Louis University School of Medicine, St. Louis, MO, U.S.A. Phone: +1 314 268 5380. Fax: +1 314 268 5736. Email: chibnajt@slu.edu Received and accepted 20 Feb 2004.

Background

Upon assuming the Editorship of *International Psychogeriatrics* (*IP*), David Ames asked me, in my role as statistical advisor to the Editor-in-Chief, to perform a "statistical audit" of one year of original research articles in *IP*. We chose 2003 (volume 15, issues 1–4). The purpose of the audit was to critically examine the statistical requirements, as dictated by *IP*, for articles published in the journal. In some ways, this audit represents an extension of a statistical review that I wrote several years ago for then-Editor Robin Eastwood (Chibnall, 2000). That review presented some fundamental statistical concepts that may not be specifically addressed in the publication of clinical research. Those comments are still relevant today, both in general and with regard to the audit, so a review of the previous paper may help to clarify the slant of this one.

Method

The method for auditing the 2003 articles was straightforward. First, only research articles that used inferential statistical techniques (as opposed to descriptive statistics) were reviewed. This excluded case-reports and descriptive studies. Second, only the statistical part of the articles was evaluated. The theoretical and methodological/design aspects of the articles were not assessed, nor were ancillary issues like sufficiency of the literature review/references, writing style, uniqueness of the investigation, extent to which conclusions were supported by the data, etc. The reviewers of the articles, of course, had already evaluated these issues. Third, articles were evaluated according to whether they addressed or met certain fundamental statistical issues in the analysis and reporting of data, as follows:

- *Did the authors report or refer to a power analysis for the study? Did the authors address issues of power in reporting results, particularly for null findings?*

This criterion is important because every study—by virtue of sample size, variability, and effect size—has more or less ability to detect significant (conventionally, $p < 0.05$) differences. Readers need to know the size of the effect that the study is capable of interpreting as "statistically significant." Either too many or too few subjects can lead to misinterpretation of the relevance of the results. Power is particularly important when studies fail to find hypothesized differences. If power is too low, the probability of a Type II error (false negative conclusion) is high. Small sample sizes can cause the analysis to miss clinically significant effects.

- *Did the authors report an effect size indicator for the results, like a confidence interval, eta^2, omega^2, Cohen's d, Cramer's V, or odds ratio?*

https://doi.org/10.1017/S1041610204000821 Published online by Cambridge University Press

This criterion speaks to the magnitude of the "statistically significant" results. As detailed in the previous report (Chibnall, 2000), a low *p*-value does not by itself confer meaning to any result. For any effect size, no matter how small, there is a sample size that will generate a $p < 0.05$ outcome. The *p*-value is a probability statement that must be evaluated across many studies. By itself, it tells us almost nothing about the magnitude of the reported results. Effect size indicators like those listed above give the reader this additional information. Sometimes a "statistically significant" finding is inconsequential, because the sample size is large enough to detect very small effects. But the reader needs to know whether the finding is clinically relevant in addition to whether it is statistically significant. This issue is not independent of power analysis. Studies should be powered sufficiently to detect the "minimum meaningful difference" for a given hypothesis (too many subjects maximizes statistical significance, but perhaps for negligible effects; too few subjects increases the probability of a Type II error). That difference should be specified ahead of time, so that if statistically significant effects are found, the reader knows immediately that the effect is large enough to be relevant. By calculating and reporting effect size indicators, the reader can evaluate the magnitude of a given effect, both absolutely and relative to other studies of the same hypotheses. Even for studies without power analyses, *post hoc* effect size indicators are useful to the reader trying to evaluate the magnitude of a given finding.

- *Did the authors do multiple univariate statistical tests without correction for Type I error inflation?*

When numerous univariate analyses are used to evaluate statistical significance among a group of correlated outcome variables, Type I error (false positive conclusion) probability is increased. Thus, accurate *p*-values depend on attention to this common problem in clinical research.

- *Did the authors use the correct statistic for the data being analyzed? Did they use the correct statistic incorrectly in the analysis?*

Sometimes the wrong statistic is used, given the type of data being analyzed. For example, ordinal or multi-categorical data are analyzed with parametric statistics (rather than nonparametric statistics). Sometimes, the correct statistic is applied incorrectly. For example, analysis of variance is used to analyze differences between three or more groups, but *post hoc* comparisons are not made.

- *Did the authors show that the data they analyzed were amenable to the type of statistical analysis they chose?*

This is a common oversight. All statistics have assumptions that must be met if *p*-values are to be accurate. This is especially true for complex multivariate statistical analyses like multivariate analysis of variance, multiple regression, logistic regression, and discriminant function analysis. It is important to

https://doi.org/10.1017/S1041610204000821 Published online by Cambridge University Press

demonstrate that the data being analyzed are not inordinately skewed, invariant, low in frequency, unequal in variance, or highly multicollinear. This becomes increasingly important as sample sizes become smaller. Sometimes the most complex analyses are performed on samples that are grossly inadequate with regard to size for the method employed (e.g., factor analysis).

- *Did the authors provide sufficient detail when reporting statistical results so that a reader could evaluate the validity of the statistical analysis?*

A reader must be able to evaluate the statistical accuracy and "history" of reported findings. This means that important detail must be included, like indices of variance, degrees of freedom, subject-to-variable ratios, names of tests used, values of statistics (e.g. F, t, χ^2, R, beta), and the various loadings, coefficients, and indexes characteristic of regression analysis, factor analysis, and discriminant function analysis.

Note that these "fundamental" statistical criteria—selected in consultation with textbooks and published commentaries on the topic of statistical analysis—represent a subset of the universe of statistical criteria on which the articles could have been evaluated. Such are the vagaries of reviews like this one. I trust that most of you will find usefulness in most of the criteria selected. More importantly, I trust that most of you will agree that attention to these issues in articles submitted to *IP* will strengthen the quality and impact of the research published therein. Thus, the audit is offered in the spirit of improving the statistical quality of *IP*.

Results

Twenty articles were reviewed. Table 1 summarizes the audit data. With respect to the criteria for review, none of the articles addressed power of the statistical analyses; half included effect size indicators; about half adjusted statistical analyses for multiple univariate comparisons, where applicable; few attended to or demonstrated the adequacy of the data being analyzed; and few provided sufficient detail to properly evaluate the statistical analyses and reported results. Most articles did, however, use the right statistic in the right way.

Discussion

The results of the audit indicate that attention to issues of statistical power and effect size is lacking in the *IP* articles reviewed. Most of the articles in Table 1 that received a "Y/N" rating for effect size indicators received the "Y" because they reported odds ratios (or, in some cases, standardized β weights for

https://doi.org/10.1017/S1041610204000821 Published online by Cambridge University Press

Table 1. Summary of statistical audit for original research articles in *International Psychogeriatrics* for the year 2003[*]

STUDY	VOLUME (ISSUE): PAGE NUMBERS	POWER ANALYSIS? *OR* ATTENTION TO POWER ISSUE?	EFFECT SIZE INDICATOR?	ADJUSTMENT FOR MULTIPLE UNIVARIATE ANALYSES?	CORRECT STATISTIC EMPLOYED? *OR* STATISTIC EMPLOYED CORRECTLY?	SPECIFIC ATTENTION TO ADEQUACY OF DATA FOR CHOSEN ANALYSIS?	SUFFICIENT DETAIL WHEN REPORTING STATISTICAL RESULTS?
Meguro *et al.*	15(1): 9–25	N	N	Y/N	Y	N	Y/N
Stewart *et al.*	15(1): 27–36	N	Y/N	Y/N	Y	N	N
Gaugler *et al.*	15(1): 37–58	N	Y/N	Y/N	Y	N	N
Brooks III *et al.*	15(1): 59–67	N	N	n/a	N[*]	N	Y
Elgh *et al.*	15(2): 121–133	N	N	Y/N	Y	N	Y/N
Senanarong *et al.*	15(2): 135–148	N	N	Y/N	Y	N	N
Mastwyk *et al.*	15(2): 149–156	N	N	N	N	N	N
Peterson and Wallin	15(2): 157–170	N	N	N	Y	N	N
Tran *et al.*	15(2): 171–179	N	Y	Y	Y	N	Y
Draper *et al.*	15(2): 187–196	N	Y/N	Y	Y	N	Y/N
Robison *et al.*	15(3): 239–251	N	Y/N	Y	Y	Y/N	Y
Mui *et al.*	15(3): 253–271	N	Y/N	N	Y	N	N
Pinner and Bouman	15(3): 279–288	N	N	N	N	N	N
Marx and Cohen-Mansfield	15(3): 289–306	N	N	N	Y	N	Y
Villalpando-Berumen *et al.*	15(4): 325–336	N	Y/N	Y/N	Y	N	Y/N
Mejía *et al.*	15(4): 337–349	N	Y/N	Y/N	Y	N	Y
Strain *et al.*	15(4): 351–366	N	Y/N	Y/N	Y	N	Y
Weiner *et al.*	15(4): 367–375	N	N	N	Y	Y	Y
Marin *et al.*	15(4): 385–398	N	Y/N	n/a	n/a	Y/N	Y/N
Baiyewu *et al.*	15(4): 399–409	N	N	N	Y	N	N

[*] In the table: N = No; Y = Yes; Y/N = criterion was met for some analyses, but not for others; n/a = criterion not applicable to the study.

[*] Statistical technique was misidentified, but applied correctly.

multiple regression), which by default convey magnitude (unlike, for example, the end result of an analysis of variance, t test, or χ^2 test. The importance of the omission of power analyses and effect sizes in *IP* articles should not be underestimated. Often, a very small p-value is enough to convey to the reader a sense of "importance" to the finding; conversely, a large p-value is enough to convey to the reader that the hypothesis was not supported. Yet, as the Methods section above indicated, this can be a mistake. Large sample sizes may generate small p-values for inconsequential effects; small sample sizes may generate large p-values for substantial effects. A power analysis forces the research to consider what magnitude of effect represents a "minimum meaningful difference," what magnitude of effect to expect from the study, and, for various sample sizes, what magnitude of effect the analysis is capable of finding "statistically significant." In combination with the actual effect sizes reported in the paper, attention to issues of power and effect magnitude is indispensable for evaluating the relevance and impact of the findings reported. For example, Weiner *et al.* (vol 15, issue 4, pp. 367–375) reported a χ^2 analysis to compare prevalence of current alcohol use between Native Americans and Whites. They concluded that "Native Americans' current use of alcohol and exposure to surgery with general anesthesia were significantly lower than Whites'." They reported a χ^2 value for the alcohol variable of 19.829 with a p-value of $< .001$. This may seem impressive, but if the effect size is computed, one finds that the Cramer's V (fourfold point correlation) value for this result is 0.17. Since V is interpreted like a correlation coefficient, it is apparent (from all effect size conventions) that this is a weak effect: race (Native American vs. White) explains less than 3% ($0.17^2 = 0.029$) of the "variance" in current alcohol use.

The problem of multiple univariate analyses, though widespread, can be fixed in relatively easy ways. First, the number of outcome variables and covariates can be reduced to the most theoretically or clinically meaningful subset. This will reduce the number of analyses and also make it easier to adopt the next suggestion: multivariate analyses should be used where possible, at least as a precursor to univariate analyses. If the number of variables is limited, it is not much harder to do a multivariate analysis than it is to do many univariate analyses. Lastly, the easiest way to control for Type I error inflation is to adopt a more conservative p-value for significance. The simplest method, Bonferroni, requires dividing the acceptable alpha level (in practice, almost always 0.05) by the number of comparisons made. Other methods are also available (e.g., the Sidak adjustment). In addition to generating more accurate p-values, reducing the number of analyses also streamlines the results and tables and makes the conclusions more obvious. One example: Senanarong et al. (vol 15, issue 2, pp. 135–148) report 29 separate p-values (t tests, χ^2 tests, Pearson correlation

https://doi.org/10.1017/S1041610204000821 Published online by Cambridge University Press

coefficients) in three separate tables for what are undoubtedly correlated variables.

The last three criteria—choice of statistic, attention to data adequacy, and sufficient statistical detail—will be addressed together. For the most part, the articles reviewed incorporated the correct statistic (with the exception of the general tendency to disregard multivariate analyses). The articles by Mastwyk *et al.* (vol 15, issue 2, pp. 149–156) and Pinner and Bouman (vol 15, issue 3, pp. 279–288) received a "No" for this criterion because they did not report any statistical analyses where statistical analysis was probably warranted. With respect to attention to the adequacy of data, few articles specifically addressed this issue. Tabachnik and Fidell (1996), in their excellent book on multivariate statistical analysis, devote an entire chapter to this issue. They discuss missing data, range and variability of scores, outliers, linearity and normality of distributions, homogeneity of variance, and multicollinearity as examples of oft-overlooked factors in data analysis that can attenuate/inflate correlations, inflate standard errors, attenuate/inflate effect sizes, generate inaccurate p-values, and produce generally unstable, nonreplicable results. One example: Gaugler *et al.* (vol 15, issue 1, pp. 37–58) report means and standard deviations (*SD*s) for hours of caregiving for activities of daily living (ADLs). For the control group, the mean ADL hours is 1.74 and the *SD* is 2.37. Thus, the SD is nearly 37% larger than the mean value. This is often indicative of a seriously skewed data distribution or one or more dramatic outliers, which can attenuate effects. This issue is compounded by their calculation of change scores, which only augments the error. The final issue is equally important to the interpretation of any published article. The statistical information should be detailed enough to allow for a sophisticated review of the reported results. This level of detail is sometimes lacking in *IP* articles (perhaps for space reasons, if nothing else). While frequencies and measures of central tendency are almost always reported, the number of articles that omit indicators of variability is far too high. Further, detailed information on sample sizes is omitted, so that one is sometimes left wondering on which sample or subsample the analysis was done. Finally, the values of statistics, degrees of freedom, factor and discriminant function coefficients, names and values of *post hoc* tests, etc., should be provided in more detail, so that there is no mystery as to the origin of that "$p < 0.001$" in a table.

As a final comment, more often than not, authors of the papers reviewed here pointed out in the Discussion section one or more of the statistical limitations of the research. This took the form of noting unadjusted multiple comparisons or small sample sizes, for example. Nevertheless, the current audit suggests that these issues and other statistical basics could be addressed more proactively in the manuscripts, particularly with respect to power, effect size, and data description.

https://doi.org/10.1017/S1041610204000821 Published online by Cambridge University Press

Conclusions

IP publishes excellent and important research in the field of geriatrics. The statistical quality of *IP* can be improved, however, by attention to a few relatively fundamental issues. Hopefully, this review highlights areas where improvement is possible, with the goal of making *IP* an even better outlet for psychogeriatric research.

Acknowledgement

The author wishes to thank Raymond C. Tait for his helpful comments on this manuscript.

References

Chibnall, J. T. (2000). Some basic issues for clinicians concerning things statistical. *International Psychogeriatrics*, 12, 3–7.

Tabachnik, B. G. and Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd ed.). New York: Harpercollins College Publishers.

https://doi.org/10.1017/S1041610204000821 Published online by Cambridge University Press