DOI: 10.1111/eci.12833

ORIGINAL ARTICLE

Diagnosis checking of statistical analysis in RCTs indexed in PubMed

Paul H. Lee* and Andy C. Y. Tse†

*School of Nursing, Hong Kong Polytechnic University, Kowloon, Hong Kong,
†Department of Health and Physical Education, The Education University of Hong Kong, New Territories, Hong Kong

ABSTRACT

Background Statistical analysis is essential for reporting of the results of randomized controlled trials (RCTs), as well as evaluating their effectiveness. However, the validity of a statistical analysis also depends on whether the assumptions of that analysis are valid.

Objective To review all RCTs published in journals indexed in PubMed during December 2014 to provide a complete picture of how RCTs handle assumptions of statistical analysis.

Methods We reviewed all RCTs published in December 2014 that appeared in journals indexed in PubMed using the Cochrane highly sensitive search strategy. The 2014 impact factors of the journals were used as proxies for their quality. The type of statistical analysis used and whether the assumptions of the analysis were tested were reviewed.

Results In total, 451 papers were included. Of the 278 papers that reported a crude analysis for the primary outcomes, 31 (27·2%) reported whether the outcome was normally distributed. Of the 172 papers that reported an adjusted analysis for the primary outcomes, diagnosis checking was rarely conducted, with only 20%, 8·6% and 7% checked for generalized linear model, Cox proportional hazard model and multilevel model, respectively. Study characteristics (study type, drug trial, funding sources, journal type and endorsement of CONSORT guidelines) were not associated with the reporting of diagnosis checking.

Conclusion The diagnosis of statistical analyses in RCTs published in PubMed-indexed journals was usually absent. Journals should provide guidelines about the reporting of a diagnosis of assumptions.

Keywords Assumption, diagnosis, protocol, statistics, trials.

Eur J Clin Invest 2017; 47 (11): 847–852

Introduction

Statistical analysis is essential for reporting the results of randomized controlled trials (RCTs) and for evaluating their effectiveness. A review found an increase in the use of statistical analyses in original articles published in the *New England Journal of Medicine*, from 73% of published papers in 1978-1979 to 87% in 2004-2005 [1].

The Consolidated Standards of Reporting Trials (CONSORT) statement recommends that all RCT reports state the statistical methods used to compare groups for all outcomes, how missing outcomes and nonadherence were handled, and the methods and rationale for all additional analysis (for example, any subgroup or adjusted analyses) conducted should also be reported [2]. However, the validity of a statistical analysis also depends on whether the assumptions of the methods were valid [3,4]. For instance, an independent sample *t*-test assumes that both samples under comparison are from populations that follow normal distributions, and a chi-square test assumes that at least 90% of the cells have an expected count greater than five [3,4]. Violations of the assumptions can lead to inflated type I and/or type II errors, biased estimation of the effect sizes and inaccurate confidence intervals.

It is important to test the assumptions of all statistical analysis used and to confirm that none are violated. However, most of the published papers in medical fields did not report such tests [3, 4]. A review showed that 33% of RCTs published in top ten Indian medical journals did not include diagnosis checking of the assumptions of the regression analysis [5]. To provide a more complete picture of how RCT reports handle assumptions of their statistical analyses, we reviewed all RCTs published in December 2014 in journals indexed in PubMed. The statistical

analysis methods used in all the included RCTs, along with the presence of diagnosis checking regarding the assumptions of these methods, were examined.

Methods

Search strategy

We used the Cochrane highly sensitive search strategy (phase 1) [6] to query PubMed for papers reporting randomized controlled trials published during December 2014 and indexed by 30 April 2015. The search was conducted to evaluate the quality of RCT reports published in 2000 [7], 2006 [8] and 2014 [9]. Two reviewers (PHL and ACYT) independently screened the abstracts and full texts to determine eligibility. Details of the search can be found elsewhere [9].

Inclusion and exclusion criteria

We adopted inclusion criteria similar to those reviewing PubMed-indexed papers published in 2000 and 2006 [7,8]. An article had to satisfy the following criteria to be included in the analysis: (i) the study subjects were humans, (ii) the trial involved at least one healthcare intervention, (iii) the participants were randomly assigned into at least two study groups with different interventions and (iv) the article was published in English. Studies were excluded if they were:(i) a cost-effectiveness, diagnostic or methodological study, (ii) a secondary publication or (iii) an early phase or pilot trial, in which a statistical analysis might not have been conducted.

Search results

A total of 1 959 abstracts were identified by the search strategy, 504 full-text papers were reviewed and 451 papers were included in the analysis.

Data extraction

For each RCT report, we identified the one statistical analysis used to draw the main conclusion regarding the effectiveness of the trial on the primary outcome. Information regarding this analysis and diagnosis checking was extracted, including the method used, whether it was a parametric or nonparametric test, whether it was a crude or adjusted analysis and whether diagnosis checking of the test was conducted. Additional information regarding the RCT was also extracted, including the type of study (parallel group: each participant was randomly assigned to one of the study groups; crossover: each participant was required to participate study groups in a random sequence; or others), the 2014 impact factor of the journal, as a proxy for the quality of the journal [10,11], endorsement of the CONSORT guidelines in the author guidelines of the journal, the specialty of the journal, whether the trial was drug-related and the source of the funding (institutional, industrial, both or none).

Results

Table 1 shows rates of inclusion for diagnosis checking regarding the assumptions of the statistical analysis used in randomized controlled trials that reported a crude analysis for the primary outcomes (n = 278). Note that 5·8% of these studies did not report whether they used a parametric or nonparametric test. Among the 114 studies comparing the continuous primary outcome between two independent groups, only 27·2% (n = 31) examined whether the outcome was normally

Table 1 Inclusion of diagnosis checking about the assumptions of the statistical analysis used in randomized controlled trials reporting a crude analysis (n = 278)*

	Freq (%)
Use of parametric/nonparametric tests	
Parametric tests only	150 (54·0%)
Nonparametric tests only	75 (27·0%)
Both parametric and nonparametric tests	37 (13·3%)
Not mentioned	16 (5·8%)
Independent sample t-test/Mann–Whitney U-test used (n = 114)	
Assumptions not checked	83 (72·9%)
Normality assumption checked†	31 (27·2%)
Anderson–Darling test	1 (0·9%)
Kolmogorov–Smirnov test	16 (14·0%)
Shapiro–Wilk test	6 (5·3%)
Visualization	4 (3·5%)
Methods not mentioned	7 (6·1%)
ANOVA/Kruskal–Wallis test used (n = 73)	
Assumptions not checked	48 (65·8%)
Normality assumption checked†	25 (34·2%)
Kolmogorov–Smirnov test	8 (11·0%)
Shapiro–Wilk test	9 (12·3%)
Visualization	5 (6·8%)
Methods not mentioned	5 (6·8%)
Chi-square test/Fisher's exact test used (n = 52)	
Assumptions not checked	49 (94·2%)
Expected count assumption checked	3 (5·8%)
Log-rank test used (n = 12)	
Proportional hazard assumption not checked	12 (100·0%)

*Studies reported tests other than those shown in the table were not included due to small sample sizes.
†Multiple tests could be used.

distributed. However, among these studies, 22·6% (n = 7) did not mention how they ascertained the normality. Among the 73 studies that compared the continuous primary outcome between three or more independent groups, only 34·2% (n = 25) examined whether the outcome was normally distributed. However, among these studies, 20·0% (n = 5) of them did not mention how they ascertained the normality. Diagnosis checking of comparisons between categorical and survival outcomes was less common, with only 5·8% of these studies reported checking the expected count assumption for a chi-square test, and none checked the proportional hazard assumption of a log-rank test.

Table 2 shows the rate of inclusion of diagnosis checking about the assumptions of the statistical analysis used in randomized controlled trials that reported an adjusted analysis for the primary outcomes. A total of six studies reported tests other than those shown in the table were not included due to small sample sizes. In general, diagnosis checking was rarely conducted, with only 20%, 8·6% and 7% checked for generalized linear model, Cox proportional hazard model and multilevel model (including generalized estimating equation), respectively.

Table 2 Inclusion of diagnosis checking about the assumptions of the statistical analysis used in randomized controlled trials reporting an adjusted analysis (n = 166)*

	Freq (%)
Generalized linear model used (n = 60)	
Assumptions not checked	48 (80·0%)
Normality assumption checked†	12 (20·0%)
Kolmogorov–Smirnov test	4 (6·7%)
Shapiro–Wilk test	3 (5·0%)
Visualization	2 (3·3%)
Skewness and Kurtosis	1 (1·7%)
Methods not mentioned	3 (5·0%)
Cox proportional hazard model used (n = 35)	
Assumptions not checked	32 (91·4%)
Proportional hazard assumption checked	3 (8·6%)
Schoenfeld residuals	1 (2·9%)
Crossing of Kaplan–Meier curves	2 (5·7%)
Multilevel regression/generalized estimating equation used (n = 71)	
Assumptions not checked	66 (93·0%)
Normality assumption checked†	5 (7·0%)
Visualization	2 (2·8%)
Methods not mentioned	3 (4·2%)

*Studies reported tests other than those shown in the table (n = 6) were not included due to small sample sizes.
†Multiple tests could be used.

Table 3 shows the associations between study characteristics, journal indicators and the inclusion of diagnosis checking. Among trials that reported an adjusted analysis for the primary outcomes (n = 172), a smaller proportion of diagnosis checking was reported among journals that have an impact factor of 3 or less (5·9%). All study characteristics, and other journal indicators, were not associated with diagnosis checking.

Discussion

It is widely accepted that statistical tests are predicated on a number of assumptions and that checking these assumptions is important, as has been documented when the tests were developed [12–14]. We found that the reporting of diagnosis checking of statistical analysis in RCTs published in PubMed-indexed journals was very poor. Even for basic statistical analyses such as *t*-tests or chi-square tests, less than 30% of the included papers reported a diagnosis of their assumptions. For more advanced statistical analyses, such as log-rank test and regressions, less than 10% of the included papers reported a diagnosis of their assumptions. Our findings were comparable to those of a small-scale study using papers published by Suez Canal University researchers, which noted that only 12% and 25% of them included a diagnosis check of the assumptions for *t*-test and regression analyses, respectively [15]. Similar results were also found in the field of social sciences, where only 11% of the papers examined the normality of the data when a univariate ANOVA test was used, and 15·5% examined normality when a repeated-measures analysis was used [16]. Our findings raise questions regarding the conclusions of most of the RCTs indexed in PubMed, as the appropriateness of their statistical tests could not be confirmed.

Most diagnosis checking tests quantify the deviation between the observed data and the expected distribution when the data satisfy the assumption perfectly. For instance, the Kolmogorov–Smirnov test computes the deviation of the empirical distribution function of the observed data with the normal cumulative distribution function, and the Shapiro–Wilk test computes the deviation of the (weighted sum of the) order statistics of the observed data with that of a normally distributed random variable. However, some of the diagnosis checking tests are very sensitive to small departures from normality especially when sample sizes are large [17], and a P-value < 0·05 of these normality tests may not invalidate statistical tests that assume normality. Researchers should use multiple methods (e.g. diagnosis checking test plus visualization) to confirm the assumption of the statistical analysis used.

We found that the quality of the journal and study characteristics, except 2014 impact factor, were not associated with the

Table 3 Inclusion of diagnosis checking about the assumptions of the statistical analysis used in randomized controlled trials by study characteristics and journal indicators (n = 262)

	Reporting crude analysis (n = 278)		Reporting adjusted analysis (n = 172)	
	Not checked (Freq, %)	Checked (Freq, %)	Not Checked (Freq, %)	Checked (Freq, %)
Study type				
Parallel	193 (80·4%)	47 (21·6%)	131 (87·3%)	19 (12·9%)
Crossover	13 (61·9%)	8 (38·1%)	10 (100·0%)	0 (0·0%)
Others	12 (70·6%)	5 (29·4%)	11 (91·7%)	1 (8·3%)
P-value	0·10		0·75†	
Drug trial				
Yes	93 (82·3%)	20 (17·7%)	70 (90·9%)	7 (9·1%)
No	125 (75·8%)	40 (24·2%)	82 (86·3%)	13 (13·7%)
P-value	0·19		0·35	
Industrial funding				
Yes	47 (79·7%)	12 (20·3%)	57 (89·1%)	7 (10·9%)
No	171 (78·1%)	48 (21·9%)	95 (88·0%)	13 (12·0%)
P-value	0·79		0·83	
Institutional funding				
Yes	106 (79·7%)	27 (20·3%)	105 (86·8%)	16 (13·2%)
No	112 (77·2%)	33 (22·8%)	47 (92·2%)	4 (7·8%)
P-value	0·62		0·43†	
Journal type				
General medical	15 (83·3%)	3 (16·7%)	26 (92·9%)	2 (7·1%)
Specialty	203 (78·1%)	57 (21·9%)	126 (87·5%)	18 (12·5%)
P-value	0·60		0·54†	
CONSORT guidelines				
Endorsed	121 (77·6%)	35 (22·4%)	83 (89·2%)	10 (10·8%)
Not endorsed	94 (79·0%)	25 (22·4%)	83 (89·2%)	10 (10·8%)
P-value	0·78		0·68	
2014 impact factor				
Not indexed	32 (80·0%)	8 (20·0%)	3 (37·5%)	5 (62·5%)
0·001-3	88 (75·9%)	28 (24·1%)	32 (94·1%)	2 (5·9%)
3·001-5	44 (78·6%)	12 (21·4%)	38 (90·5%)	4 (9·5%)
5·001-10	37 (78·7%)	10 (21·3%)	38 (86·4%)	6 (13·6%)
>10	17 (89·5%)	2 (10·5%)	41 (93·2%)	3 (6·8%)
P-value	0·76		0·003†	
Total	218 (78·4%)	60 (21·6%)	152 (88·4%)	20 (11·6%)

†Fisher's exact test.

reporting of diagnosis checking. To our surprise, studies published in journals not indexed by Journal Citation Reports (JCR) were actually more likely to report a diagnosis of an adjusted statistical analysis. This contradicts the finding that report quality increased with the impact factor of the journal [9, 18]. One possible explanation is that journals with high impact factors usually impose length limits on manuscripts. As a result, authors submitting their manuscripts to these journals did not have enough space to report the details of diagnosis of assumptions.

However, we suspect that most authors of the reviewed RCT reports did not even know the assumptions of the statistical analyses they were using, and if they did, they apparently did not know how these assumptions were to be examined. In a survey of 30 PhD students, over 85% were unfamiliar with the assumptions of *t*-tests, ANOVA tests and regressions, while 60% were unfamiliar with how to check these assumptions [19]. In addition, many researchers misunderstood the concept that the Central Limit Theorem suggests normality when a sample size is larger than 30 [20]. Therefore, to improve the statistical analysis of RCTs, we recommend providing guidelines to authors about how assumptions of statistical tests should be diagnosed, and to introduce remedies if these assumptions are violated. For example, the Statistical Analyses and Methods in the Published Literature (SAMPL) published by the European Association of Science Editors [21] provide a number of recommendations to the authors to report the statistical methods and results of their studies. In particular, concerning the assumptions of the statistical analysis, the SAMPL guidelines suggested the authors to verify three important aspects of the data, including skewness (that skewed data were analysed with appropriate nonparametric methods), paired data (that analysed using paired methods) and linearity (that linear regression should only be used when the underlying associations are linear). In addition, the online journal *Frontiers of Psychology* has recently published a series of articles on testing assumptions [22]. Teaching materials about the assumptions of statistical analyses exist at undergraduate level [20,23], and further studies testing the effectiveness of these teaching materials are warranted. Note that the diagnosis of assumptions with the observed data is not the best approach, as the data are only the realization of the underlying study population data, in which the assumptions of the statistical analysis actually rely on [24]. Unfortunately, the population data are unobservable, and some researchers advocate using established prior knowledge and empirical evidence to determine the validity of the statistical assumptions in addition to the observed data [24], or using statistical analysis methods that are robust to the assumptions [25].

Our study is not without limitations. We were only able to assess whether the assumptions were reported, but not whether they were checked. More importantly, we were unable to assess the impact to the statistical analysis of diagnosis checking. Of the 504 papers reviewed, 12 were excluded because the full texts were not accessible to us. We believe this was a minor limitation as it represented just 2·6% of the reviewed papers. We searched only one database, so RCT studies not published in journals indexed in PubMed were not included.

To conclude, the diagnosis of statistical analyses in RCTs published in PubMed-indexed journals was often lacking. Journals should provide guidelines about the reporting of the diagnosis of assumptions.

Funding

None.

Author Contributions

Paul H. Lee drafted the manuscript and conducted the data analysis. Paul H. Lee and Andy C. Y. Tse conducted the systematic review. Andy C. Y. Tse critically reviewed the manuscript.

Address

School of Nursing, Hong Kong Polytechnic University Kowloon, Hong Kong (Paul H. Lee); Department of Health and Physical Education, the Education University of Hong Kong New Territories, Hong Kong (Andy C. Y. Tse).

Conflict of interest

None declared.

Correspondence to: Dr Paul H. Lee, School of Nursing, GH527, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Tel.: +852-3400 8275; fax: +852-2364 9663; email: paul.h.lee@polyu.edu.hk

Received 18 May 2017; accepted 28 August 2017

References

1 Horton NJ, Switzer SS. Statistical methods in the Journal. *New Engl J Med* 2005;**353**:1977–9.
2 Schulz KF, Altman DG, Moher D and for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Lancet* 2010;**375**:1136.
3 Lang T. Twenty statistical errors even you can find in biomedical research articles. *Croat Med J* 2004;**45**:361–70.
4 Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research–a review of common pitfalls. *Swiss Med Wkly* 2007;**137**:44–9.
5 Hassan S, Yellur R, Subramani P, Adiga P, Gokhale M, Iyer MS *et al*. Research design and statistical methods in Indian medical journals: a retrospective survey. *PLoS ONE* 2015;**10**:e0121268.
6 Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol* 2002;**31**:150–3.

7 Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005;**365**:1159–62.
8 Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ* 2010;**340**:c273.
9 Lee PH, Tse ACY. The quality of the reported sample size calculations in randomized controlled trials indexed in PubMed. *Eur J Intern Med* 2017;**40**:16–21.
10 Garfield E. Journal impact factor: a brief review. *CMAJ* 1999;**161**:979–80.
11 Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. *JAMA* 2002;**287**:2805–8.
12 Pearson E. The analysis of variance in cases of non-normal variation. *Biometrika* 1931;**23**:114.
13 Pearson K. Mathematical contribution to the theory of evolution. VII: on the correlation of characters not quantitatively measurable. *Philos Trans R Soc Lond B Biol Sci* 1901;**195**:1–47.
14 Student. The probable error of a mean. *Biometrika* 1908;**6**:1–25.
15 Nour-Eldein H. Statistical methods and errors in family medicine articles between 2010 and 2014-Suez Canal University, Egypt: a cross-sectional study. *J Family Med Prim Care* 2016;**5**:24–33.
16 Keselman HJ, Huberty CJ, Lix LM, Olejnk S, Cribbie RA, Donahue B *et al*. Statistical practices of educational researchers: an analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Rev Educ Res* 1998;**68**:350–86.
17 Peat J, Barton B. Medical Statistics: A Guide to Data Analysis and Critical Appraisal. Oxford: Blackwell; 2005.
18 Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009;**338**:b1732.
19 Hoekstra R, Kiers HAL, Johnson A. Are assumptions of well-known statistical techniques checked, and why (not)? *Front Psychol* 2012;**3**:137.
20 Cummiskey K, Kuiper S, Sturdivant R. Using classroom data to teach students about data cleaning and testing assumptions. *Front Psychol* 2012;**3**:354.
21 Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A, editors. *Science Editors' Handbook*. European Association of Science Editors; 2013: 228pp. Available at: https://www.ease.org.uk/publications/science-editors-handbook/. Accessed on 23 July 2017.
22 Osborne JW. Is data cleaning and the testing of assumptions relevant in the 21st century? *Front Psychol* 2013;**4**:370.
23 Nimon KF. Statistical assumptions of substantive analyses across the general linear model: a mini-review. *Front Psychol* 2012;**3**:322.
24 Wells CS, Hintze JM. Dealing With Assumptions Underlying statistical tests. *Psych School* 2007;**44**:495–502.
25 García-Pérez MA. Statistical conclusion validity: some common threats and simple remedies. *Front Psychol* 2012;**3**:325.

