Established in 1871

Swiss Medical Weekly

Formerly: Schweizerische Medizinische Wochenschrift

An open access, online journal • www.smw.ch

Review article: Current opinion | Published 4 February 2015, doi:10.4414/smw.2015.14076

Cite this as: Swiss Med Wkly. 2015;145:w14076

Statistical analysis and reporting: common errors found during peer review and how to avoid them

Gillian Worthy

Kleijnen Systematic Reviews, York, United Kingdom

Summary

When performing statistical peer review for Swiss Medical Weekly papers there often appear to be common errors or recurring themes regarding the reporting of study designs, statistical analysis methods, results and their interpretation. In order to help authors with choosing and describing the most appropriate analysis methods and reporting their results, we have created a guide to the most common issues and how to avoid them. This guide will follow the recommended structure for original papers as provided in the guidelines for authors (http://blog.smw.ch/what-smw-has-to-offer/guidelines-for-authors/), and provide advice for each section. This paper is intended to provide an overview of statistical methods and tips for writing your paper; it is not a comprehensive review of all statistical methods. Guidance is provided about the choice of statistical methods for different situations and types of data, how to report the methods, present figures and tables, and how to correctly present and interpret the results.

Key words*: statistics; analysis methods; reporting; tables; figures*

Introduction

When performing statistical peer review for Swiss Medical Weekly papers there often appear to be common errors or recurring themes regarding the reporting of study designs, statistical analysis methods, results and their interpretation. In order to help authors with choosing and describing the most appropriate analysis methods and reporting their results, we have created a guide to the most common issues and how to avoid them. This is not intended to provide advice on study design; once a study has been completed and the paper submitted for peer review the design cannot be altered. Good statistical analysis cannot benefit a poorly designed study and it is recommended that assistance in designing the study is sought from a statistician. An excellent textbook on study design that covers the design of, and sample size calculations for different study designs, including randomised controlled trials, cross-sectional, cohort and case-control studies, as well as surveys is provided by Machin and Campbell [1]. Not all studies will require sample size calculations, for example, pilot or small-scale feasibility studies which are the first assessment of a treatment in a particular setting and are used to collect data to inform the design of a larger study. However, sample size calculations should be undertaken for a randomised controlled trial to ensure that it has sufficient statistical power to detect an effect in the primary outcome of interest. An introduction to sample size calculations is provided by Noordzij et al. [2].

This guide will follow the organisation for original papers as provided in the guidelines for authors (http://blog.smw.ch/what-smw-has-to-offer/guidelines-for-authors/), and provide advice for each section. Authors should make sure that they provide a clear statement of the study design and ensure that their reporting follows the recommended reporting guidelines for that design, as provided by the EQUATOR network (http://www.equator-network.org/). Other papers and text books providing guidance on statistical analysis and reporting are available [3–7] including previous guides published in Swiss Medical Weekly [8–10].

Summary

Ensure the results reported in the summary are consistent with those in the main text.

Do not report additional results which have not appeared in the main text.

Introduction

Please provide a clear aim. A common problem is that the aim of the study is not very clear, or appears to differ from the aim addressed by the results and discussion. Use the PICOS framework as a guide, which covers: P population under evaluation; I intervention(s) being assessed; C comparators; O outcomes; S study design.

Also state why there is a need for your study; maybe there is a lack of research in a particular area, or a clear need for additional evidence. Make sure your research is original and not repeating previous work.

Material and methods

– If possible report the hypotheses under evaluation in the analysis. If there were no pre-specified hypotheses and the analysis is exploratory then make this clear; data dredging should be avoided.
– Outcomes: provide a separate section detailing all the study outcomes, how they were measured, when and by whom (as appropriate). Split it into primary and secondary outcomes if relevant, especially for a clinical trial. All outcomes need to be listed to prevent outcome reporting bias (only reporting those outcomes which show statistically significant or favourable results).
– Details of the patients such as the number included in the study, age and gender are results, not methods and should be part of the description of the data in the first part of the results section.

Statistical methods

The statistical methods section is often poorly reported. Details of all statistical tests and models should be reported in sufficient detail to enable the reader to understand what has been done. All analysis methods should be reported, the outcomes being analysed and which comparisons are being made. Details of how the results are reported should also be given. For example, quality of life data are summarised using means and standard deviations, results from logistic regression models are reported as odds ratios with 95% confidence intervals (CI).

All analyses listed in the methods should have a corresponding set of results and *vice versa*, it is quite common to find results being reported which have not been previously mentioned in the methods section. The number of statistical tests or analyses should be kept to a minimum and ideally pre-specified in order to avoid multiple hypothesis testing. I did once review a paper that had more statistical tests than participants!

This section is split into tips regarding the choice of analysis method, and how to report them.

Choosing an appropriate statistical analysis method

A summary of the statistical analysis methods applicable to continuous and categorical data and different numbers of groups is presented in table 1 (adapted from Petrie [11]). Other issues are discussed below, this is not intended to be a complete list, but covers the main points arising from the statistical review of recent submissions. Before performing any statistical analysis it is important to summarise the data, and assess any underlying assumptions required by the statistical tests.

Descriptive statistics

Descriptive statistics should be used to summarise the data, especially the characteristics of the study population. Continuous data should be summarised using means and standard deviations (SD) for normally distributed variables, or medians and ranges (or inter-quartile ranges) if the variable is skewed. Categorical data should be summarised using numbers and percentages.

Parametric versus non-parametric tests

It is the test which is parametric or non-parametric NOT the data. Statements such as 'Non-parametric data are presented as median and range' are incorrect. Analysis methods such as a t test require that the data follow a normal distribution. If this assumption is doubtful then transforming the data (e.g., by taking logarithms) can often help. If data transformation does not improve the distribution or is not appropriate, then use the relevant non-parametric test (see table 1) although note that these have less statistical power (are less likely to detect a true effect).

Correlation and regression

Correlation measures the degree of linear association between two numerical variables, not agreement or 'cause and effect'. For assessing whether one or more variables can predict another regression is needed, correlation and regression are often confused. Correlation analyses should be accompanied by scatterplots so the reader can visualise the patterns of the data and whether there are any outlying values. There are different methods for calculating the correlation coefficient, the two most common are: Pearson (assumes that at least one of the two variables is normally distributed) and Spearman (the non-parametric equivalent which can be used for smaller samples, where one or both are ordinal variables, or when the relationship is non-linear).

Categorising continuous variables

This is often done and should be avoided as it reduces statistical power. The choice of cut-off points could influence the results, especially if they were chosen once data analysis had started. Unless an acceptable clinical categorisation (such as cholesterol lowering thresholds) is being used, continuous variables should be left as they are in regression modelling.

Paired or clustered data

If two measurements are made on each participant such as before and after treatment then it is incorrect to treat these as two separate measurements as the within patient correlation needs to be accounted for. Paired data needs to be analysed with paired tests (see table 1). Clustered data, including repeated measurements over time (such as quality of life) also need to be analysed using methods which account for the fact that there were multiple measurements on the same participant. Options include using a simple summary measure (overall mean, change from baseline to a specified time, the maximum value, or the area under the curve over the whole time period); repeated measures regression; or more complex regression models (multilevel models, generalised estimating equations).

Multivariable regression

Multiple or multivariable regression seems to be less widely used in papers and the peer review process often suggest that this is included in a paper. Multivariable regression should be used to adjust for any variables that

differ between groups in an observational study, to adjust treatment estimates in a randomised controlled trial for any known prognostic factors, or to look at the effect of a variable when accounting for the effects of other variables (e.g., age and gender). Specifically analyses of mean change or percentage change from baseline need to adjust for each participant's baseline value (for example reduction in wound area). However, the size of the study needs to be considered in that a multivariable regression would require more data than a simple univariable regression (which contains only one variable). Approximately 10 people with the outcome need to be included for each variable in the model, so an analysis of blood pressure adjusting for age, gender and baseline blood pressure would need to include at least 30 people.

A continuous outcome should be analysed with linear regression, counts or rates with Poisson regression, categorical outcomes with logistic regression and time to event outcomes with Cox proportional hazards regression or a parametric survival model (see below). A helpful guide to the methods and interpretation of multivariable analyses is given by Katz [12].

Survival analysis

Time to event data, such as time to healing or progression-free survival should be analysed using appropriate survival analysis methods. Using the mean time to event for those who experienced the event is incorrect as this loses information about those who were lost to follow-up or did not experience an event. Survival curves should be plotted and survival can be compared between groups using a log-rank or Wilcoxon test. Regression models such as the Cox proportional hazards model (the underlying proportional hazards assumption should be checked) or parametric models (such as Weibull) can be used to adjust for other variables.

Diagnostic tests

The performance of a diagnostic test or measurement should be compared to a reference or gold standard test or measurement. Ideally all participants should undergo both tests. For a binary outcome (diseased or not diseased) a 2 by 2 table should be presented, from which measures of sensitivity, specificity, positive and negative predictive values with 95% CI can be calculated. For a continuous test score a receiver operating characteristic (ROC) curve can be used and the area under the curve with 95% CI calculated. If one or more cut-off thresholds have been used to calculated sensitivity or specificity these should be clearly reported along with the reasons for their choice.

Reporting analysis methods

It should be clear from the description which variables were analysed with each different analysis method. Vague statements such as 'data were analysed with the chi-squared test, t-test and regression' are not helpful, as it is unclear which data were analysed with each method.

– If there was a sample size calculation then report it in sufficient detail to enable it to be replicated by a statistician. This requires information about the type I error (alpha, usually 0.05), type II error (1 – beta, the power often 80% to 90%), the minimum clinically relevant difference (the smallest difference between the groups that would be clinically relevant), and the outcome for the control group based on previous research (the event rate for a dichotomous outcome, or the mean and SD for a continuous outcome).
– If there was no sample size calculation but there was some information about the study size then do report this ('no formal sample size calculation was performed but all available patients in two centres were included in the study', or 'this was a pilot study and a sample size calculation was not relevant').
– Report full details of how the underlying analysis assumptions were checked (e.g. normal distribution, constant variance between groups, and a linear relationship between two variables for correlation or regression) and how any transformations were performed.
– Analyses should, where possible, be accompanied by relevant plots. Scatterplots for correlation, survival curves for time-to-event analyses, boxplots or means with 95% CI for summaries of continuous variables,

Table 1: Choosing the correct statistical test.

Number of groups	Continuous outcomes	Categorical outcomes
1	One-sample t test	Test of a single proportion (based on estimated proportion and its standard deviation)
	Sign test (non-parametric*)	Sign test (non-parametric*)
2	Two-sample t-test	Chi-squared test (or Fisher's exact test if any values are <5)
	Wilcoxon rank sum/Mann-Whitney U test (non-parametric*)	McNemar's test (paired data)
	Paired t-test (paired data)	Mantel-Haenszel chi-squared test (for stratified odds ratios)
	Wilcoxon signed rank test (paired data non-parametric*)	Logistic regression (for assessing the effect of one or more explanatory variables)
	Linear or multiple linear regression (for assessing the effect of one or more explanatory variables)	
3 or more	Analysis of variance (ANOVA)	Chi-squared test
	Kruskal-Wallis (non-parametric*)	Chi-squared test for trend (for ordered categories, e.g., mild, moderate, severe pain)
	Linear or multiple linear regression (for assessing the effect of one or more explanatory variables)	Logistic regression (for assessing the effect of one or more explanatory variables)

* Non-parametric indicates the equivalent non-parametric test which does not make any assumptions about the distribution of the data. T-tests and ANOVA assume that the data being analysed follow a normal distribution with similar variance in each group. This is not intended to be an exhaustive list, for details of other statistical methods consult a suitable textbook or seek advice from statistician.

ROC curves for diagnostic tests, forest plots for meta-analyses.
– Full details of the modelling methods for any multivariable analyses should be specified, including the model type(multiple linear, logistic, Cox proportional hazards), the outcome being analysed, which variables were assessed for inclusion in the model and the selection method (forwards, backwards, stepwise, etc.) and the p-values used to include or exclude variables.
– Unusual or more complex statistical methods should be referenced.
– If there was any adjustment for multiple hypothesis testing to prevent the chance of a false positive finding (e.g., applying a Bonferroni correction or using smaller p-values such as 0.01 instead of the conventional 0.05). This can also be minimised by pre-specifying the analyses and keeping them to a minimum number.
– Details of the statistical software used, whether hypothesis tests were one or two-sided (most should be two-sided unless there was a strong belief or previous evidence about the direction of the results) and the p-values used to conclude statistical significance should be reported.

Results

All analysis results should be reported, not just those which are statistically significant. If there are a lot of results they do not all need to be reported in the main text but all results should be available in tables, figures or appendices.
– Provide the start and end dates of recruitment, the number of participants recruited, and the number analysed (see the EQUATOR network guidelines for examples of participant flowcharts http://www.equator-network.org/) and a brief description of the participants.
– Please report effect sizes (mean differences, odds ratios, hazard ratios, etc.) with 95% confidence intervals (or standard errors [SE]). Other measures such as correlation coefficients and areas under curves also should be reported with 95% CI. If different CI have been reported, such as 90% or 99% please make this clear.
– Results which are just describing the data should be reported as mean and SD. Results from statistical tests or models should be reported as the effect size (see above) with the corresponding 95% CI (or SE). SD and SE are often confused. The SD is a measure of the variation in the data and the SE is a measure of the variation in the estimate from the statistical analysis. The SE is affected by the sample size, a larger dataset will provide more precise estimates of the outcome in question with narrower CI (as SE = SD/√sample size).
– For survival analyses report the median survival time with 95% CI for each group (if it was reached) alongside the p-value from a test comparing survival curves. If a regression model was used also report the hazard ratio with 95% CI.
– Report p-values in full (to 2 or 3 decimal places). Very small values such as $p < 0.001$ can be reported as such but avoid the use of *, **. Do not use 'NS', '>0.05' for results which are not statistically significant.
– For regression models report a measure of the 'goodness of fit' of the model to the data, e.g., R^2 or a Hosmer-Lemeshow test.

Tables and figures

– Please provide a table summarising participant details using descriptive statistics (mean, standard deviation, median, inter-quartile range, number and percentage). For a randomised controlled trial it is not appropriate to report p-values comparing groups at baseline.
– Make sure it is clear which statistics are being reported, either through labels in the table or as a footnote. For example, 34 (2.8) is the mean and standard error.
– Report the number of participants in each group for tables which report descriptive data. Also provide the numbers included in each analysis on all tables and/or figures which contain results. Check that percentages are correct.
– Results of regression models should be reported in full in the tables (i.e., regression coefficients and SE, or effect sizes with 95% CI or SE and p-values, for all the terms in each model).
– All figures should have clear titles.
– All figures should have clearly labelled axes with units, and any symbols should be labelled. It is quite common to see symbols on figures without any indication of what they represent.
– Do not make your figures too complicated by including too much information or too many groups.

Discussion

– Only discuss those results which have been presented in the results section. It is a common error to find extra results in the discussion which haven't previously been reported.
– Do not repeat effect sizes and confidence intervals from the results.
– Check that all results have been interpreted correctly in terms of the statistical and clinical significance and the direction of effects.

Funding / potential competing interests: No financial support and no other potential conflict of interest relevant to this article was reported.

Correspondence: *Gillian Worthy, MSc, Kleijnen Systematic Reviews Ltd., 6 Escrick Business Park, Riccall Road, UK-Esrick,York YO19 6FD, United Kingdom, gill[at]systematic-reviews.com*

References

1 Machin D, Campbell MJ. Design of studies for medical research. Chichester: Wiley; 2005.

2 Noordzij M, et al. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant. 2010;25:1388–93.

3 Altman D, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. BMJ. 1983;286:1489–93.

4 Lang T. Twenty statistical errors even YOU can find in biomedical research articles. Croat Med J. 2004;45(4):361–70.

5 Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook, European Association of Science Editors, 2013.

6 Campbell MJ, Machin D. Medical statistics a commonsense approach. 3rd ed. Chichester: Wiley; 2003.

7 Kirkwood BR, Sterne JAC. Essential medical statistics, 2nd ed. Oxford: Blackwell; 2003.

8 Young J. When should you use statistics? Swiss Med Wkly. 2005;135:337–8.

9 Young J. Statistical errors in medical research – a chronic disease? Swiss Med Wkly. 2007;137:41–3.

10 Strasak AM, et al. Statistical errors in medical research – a review of common pitfalls. Swiss Med Wkly. 2007;137:44–9.

11 Petrie A, Sabin C. Medical statistics at a glance. 3rd ed. Chichester: Wiley; 2009.

12 Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med. 2003;138:644–50.