

Review Article

The presentation of statistics

Bryant TN. The presentation of statistics.
Pediatr Allergy Immunol 1998; 9: 108–115. © Munksgaard 1998.

This article outlines the statistical requirements that authors should fulfill when submitting manuscripts to *Pediatric Allergy and Immunology*. The requirements are based on the 'Uniform Requirements for Manuscripts Submitted to Biomedical Journals' and the CONSORT statement. Common statistical flaws that routinely arise in the medical literature are described. The use of $\pm$, confidence intervals and P values, correlation and regression, multiple testing and repeated measures are discussed.

T. N. Bryant

Medical Statistics and Computing, University of Southampton/ Southampton General Hospital, Southampton, Hampshire, UK

Key-words: confidence intervals, p value, correlation regression, multiple testing, repeated measures.

T. N. Bryant, Medical Statistics and Computing, University of Southampton, South Block (Mail-point 820), Southampton General Hospital, Southampton SO16 6YD, UK

One of the dangers introduced by statistical software is that they allow the user to do what they want with their data without providing sensible warnings of the incorrect actions on their part. If a user decides to perform a parametric statistical test on data that is clearly not normally distributed the software package will oblige and produce the output without hesitation. An example of one of the few warning/help messages that is given by statistical packages is when a χ^2 calculation is carried out and a message saying "1 cell(s) had an expected frequency of less than 5" is displayed. What users should be looking for in statistical software of the future are messages that say "You should not carry out a correlation calculation using nominal data," or "you have requested a *t*-test but it is not really appropriate because the dependent variable does not seem to be normally distributed, do you want the equivalent non-parametric test performed?" However this sort of advice does not seem to be universally available. The best that we have at the moment is the flow chart approach. "Do you want to compare 2 groups? (Y/N)," "Are the data paired? (Y/N)," "Does the data follow a normal distribution (Y/N)," leading to a suggestion as to the most appropriate test. But this still assumes the user has a good understanding of the issues involved and many statisticians are concerned that this results in "a little knowledge is a dangerous thing."

This article describes the mistakes that can be made when presenting statistical information. These mistakes sometimes get into the literature only to be perpetuated by others when preparing their manuscripts. To describe all of the statistical and graphical aspects authors should consider when writing a paper would result in textbooks on medical statistics and the presentation of medical and scientific data.

It is not possible to cover all topics, or even some topics in sufficient depth, so the reader is referred to the following texts:

1. Medical Statistics - A common sense approach (1).
2. Statistics with Confidence - Confidence intervals and statistical guidelines, (2).
3. Practical Statistics for Medical Research (3).
4. The Elements of Graphing Data (4).
5. Graphing Statistics and Data - Creating better charts (5).

Describing the Statistical Methods

The journal will be adopting the Uniform Requirements for Manuscripts Submitted to Biomedical Journals (6). The International Committee of Medical Journal Editors (ICMJE) recommend the following for the statistical section of a paper:

Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information. Discuss the eligibility of experimental subjects. Give details about randomization. Describe the methods for and success of any blinding of observations. Report complications of treatment. Give numbers of observations. Report losses to observation (such as dropouts from a clinical trial). References for the design of the study and statistical methods

Table 1. Items that should be included in reports of randomised trials (reproduced from JAMA (7))

Heading	Sub-heading	Descriptor	Was it reported? Yes or No	If Yes, What Page No?
Title		Identify the study as a randomised trial.		
Abstract		Use a structured format.		
Introduction		State prospectively defined hypothesis, clinical objectives, and planned subgroup or covariate analyses.		
Methods	Protocol	Describe: Planned study population, together with inclusion or exclusion criteria. Planned interventions and their timing. Primary and secondary outcome measure(s) and the minimum important difference(s), and indicate how the target sample size was projected. Rationale and methods for statistical analyses, detailing main comparative analyses and whether they were completed on an intention to treat basis. Prospectively defined stopping rules (if warranted).		
	Assignment	Describe: Unit of randomisation (for example, individual, cluster, geographic). Method used to generate the allocation schedule. Method of allocation concealment and timing of assignment. Method to separate the generator from the executor of the assignment.		
	Masking (blinding)	Describe Mechanism (for example, capsule, tablets). Similarity of treatment characteristics (for example, appearance, taste). Allocation schedule control (location of code during trial and when broken). Evidence for successful blinding among participants, person doing intervention, outcome assessors, and data analysts.		
Results	Participant flow and follow up Analysis	Provide a trial profile (Figure) summarising participant flow, numbers and timing of randomisation assignment, interventions, and measurements of each randomised group. State estimated effect of intervention on primary and secondary outcome measures, including a point estimate and measure of precision (confidence interval). State results in absolute numbers when feasible (for example 10/20, not 50%). Present summary data and appropriate descriptive and inferential statistics in sufficient detail to permit alternative analyses and replication. Describe prognostic variables by treatment group and any attempt to adjust for them. Describe protocol deviations from the study as planned, together with the reasons.		
Discussion		State specific interpretation of study findings, including sources of bias and imprecision (internal validity) and discussion of external validity, including appropriate quantitative measures when possible. State general interpretation of the data in the light of the totality of the available evidence.		

should be to standard works when possible (with pages stated) rather than to papers in which the designs or methods were originally reported. Specify any general-use computer programs used.

Put a general description of methods in the Methods section. When data are summarized in the Results section, specify the statistical methods used to analyze them. Restrict tables and figures to those needed to explain the argument of the paper and to assess its support. Use graphs as an alternative to tables with many entries; do not duplicate data in graphs and tables. Avoid non-technical uses of technical terms in statistics, such as “random” (which implies a randomizing device), “normal,” “significant,” “correlation,” and “sample.” Define statistical terms, abbreviations, and most symbols.

When describing controlled trials, authors should conform to the CONSORT statement (7) where the methods section should contain sub-headings describing: protocol, assignment and

masking (blinding). This information is presented in Table 1. The manuscript should also contain the profile of the trial as illustrated in Figure 1. The example given in this figure is for a two group comparison; a different diagram would be required for a different study design. The report should include a justification of why the number of subjects were selected and details of the sample size calculation given. For details of sample size calculations see Machin *et al.* (8)

A statistical methods section might include:

Comparisons between controls and subjects were analysed using the Mann-Whitney test. Paired data were analysed using the Wilcoxon signed ranked test. Data were analysed using the *ABCDE* package, version *x.y*.

However if a more complex analysis is used then reference to the relevant methodology should be included.

Profile of a Randomised Controlled Trial

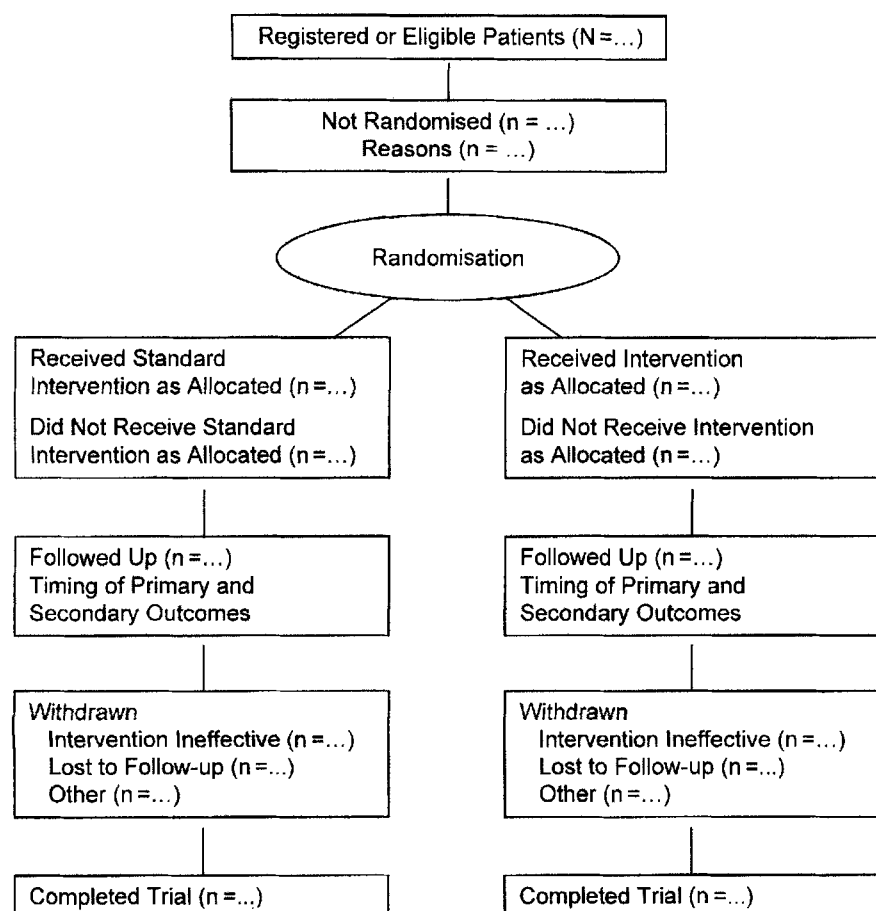


Fig. 1. Progress through the various stages of a trial, including flow of participants, withdrawals, and timing of primary and secondary outcome measures. (reproduced from JAMA (7))

Avoid using $\pm$ in the text

You often see $12.9 \pm 2.4 \mu\text{g/l}$ in the text of an article. The use of this notation is to be discouraged because, used like this, it could imply one of two parameters, either the sample standard deviation of the data (SD) or an estimate of the standard deviation of a statistical quantity, known as the standard error (SE), for example, of the mean. It is preferable to use $12.9 \mu\text{g/l}$ (SD $2.4 \mu\text{g/l}$) or $12.9 \mu\text{g/l}$ (SE $2.4 \mu\text{g/l}$), since even if the use of the symbol $\pm$ has been defined previously the reader does not have to refer back to the definition because (SD $2.4 \mu\text{g/l}$) is immediately understood.

Quoting P values

Presenting the test statistic, degrees of freedom and P value helps the reader decide if the test has been applied correctly. P values should be quoted to two decimal places and the exact P value e.g. $P = 0.02$ used. Most statistical packages display P values so there is no reason for using $P < 0.05$. If P is very

small (and therefore highly significant) $P < 0.001$ is reasonable and would be used where P is presented as 0.000 or 0.0001 by the software. Where results are considered not significant, the P value should still be given. The use of *ns* instead of $P = 0.25$ is less informative to the reader and can hide situations where significance was just missed, for example, presenting $P = ns$, rather than $P = 0.07$. This allows the reader to consider whether a larger study might have shown a difference. However, the use of P values alone is no longer considered acceptable.

Confidence Intervals or P values

For some years the leading medical journals have recommended that confidence intervals as well as P values should be used when reporting the main study results. A collection of papers on this subject were assembled in the book *Statistics with Confidence – confidence interval and statistical guidelines* (2). The editors, Gardner and Altman, argue that too much emphasis on the P values that are produced from hypothesis testing detracts from

more useful approaches, such as estimation and confidence intervals, when interpreting study results. The use of confidence intervals is recommended by the ICMJE (6).

Relying on an arbitrary P value cut off at a pre-defined level, typically 5%, leads to the wrong type of thinking. If this approach is taken, $P = 0.049$ is declared significant and $P = 0.051$ as not significant. Using a binary cut off is inappropriate and tends to lead to statistical significance being equated with clinical significance. A very small improvement, for example of 1%, of one treatment compared to another in a study may give a statistically significant ($P < 0.001$) result. However, this result alone is unlikely to change medical practice, unless other benefits, such as cost or lack of side effects, are taken into consideration. In such a case, if only P values are quoted and an estimate of the size of the difference is not given, the uncritical reader could fall into the trap of thinking that treatment A was more effective than treatment B because $P < 0.001$. This is where confidence intervals are useful because, in this example, they would give a range of values for the estimated difference and there is a 95% chance that the indicated range includes the true difference. Note that other percentages (typically 90% and 99%) are occasionally used. Confidence intervals are not required for all results. For example, we would not require the confidence interval for the mean response of subjects to treatment A and the mean response to subjects to treatment B if our major outcome was to assess the difference between treatments A and B. Only confidence intervals for the difference between treatments A and B would be required. It is not proposed that P values are not quoted, rather that they are used in context with confidence intervals, for example:

Where more than two groups are compared, reporting is more complicated and the means and confidence intervals for each group might be presented rather than the differences. For example, Carlsen et al. (9) studied inflammation markers and symptom activity in children with bronchial asthma. They compared the serum eosinophilic cationic protein (sECP) in three groups of children, symptom free and those with episodic and persistent asthma. A one way analysis of variance showed a significant difference between the three groups ($P < 0.001$) the mean values and 95% confidence intervals for sECP ($\mu\text{g/l}$) were presented as:

Group	Mean	95% CI
Symptom-free	13.5	(10.1 to 16.9)
Persistent	21.4	(16.5 to 26.4)
Episodic	29.2	(22.3 to 36.2)

Here the reader has useful information about the estimate of the mean values. We can expect that in 95 studies out of 100 the mean sECP level for symptom-free children will be between 10.1 and 16.9 and by inspecting the table it is probable that this group is different from the other two groups in the study. It is better to quote confidence intervals as (10.1 to 16.9) rather than (10.1–16.9) because the using the latter notation has disadvantages if negative values are involved.

Correlation or Regression?

The use, or abuse, of the correlation coefficient is common in many papers. Before calculating any correlation coefficient the data should be visualised by plotting a scatterplot. This will show whether we are justified in testing for a linear association between the two variables of interest because there are instances where the correlation is not an appropriate measure (1):

1. When the data is not linear (Fig. 2a). A correlation coefficient of 0.975 and a P value of < 0.001 is obtained for the following pairs of values, 1,1; 2,4; 3,9; 4,16; 5,25; 6,36; 7,49; 8,64; 9,81 and 10,100, which have a perfect association using the equation $y = x^2$ which is not linear. Similarly if the data for $y = x^2$ is used for the integers $x = -10$ to $+10$ we will obtain a correlation coefficient of 0 which falsely implies there is no association (Fig. 2b). There is a mathematical association but it is not linear, which is what correlation is testing.
2. When there are outliers in the data, in which one or two observations in the data plot away from the other observations. Correlation coefficients are sensitive to the presence of outliers which strongly influence their estimated value (compared Fig. 2c and 2d).
3. Care should be taken if more than one distinct group is included in the calculation, for example normal and atopic subjects. A scatterplot (Fig. 2c) where the subjects in each group are plotted (Fig. 2e) using different symbols should reveal such possible problems. Another example is where there are two distinct clouds of data points, each of which have zero correlation, treating these as one single group could result in a spurious significant correlation.
4. If one of the variables is determined in advance, for example drug concentration, then it is inappropriate to use correlation: regression is more appropriate. This is because the drug concentration affects the response that is measured (e.g. FEV_1) whereas the reverse cannot be the case.
5. A correlation coefficient is not appropriate when comparing a new method of measurement

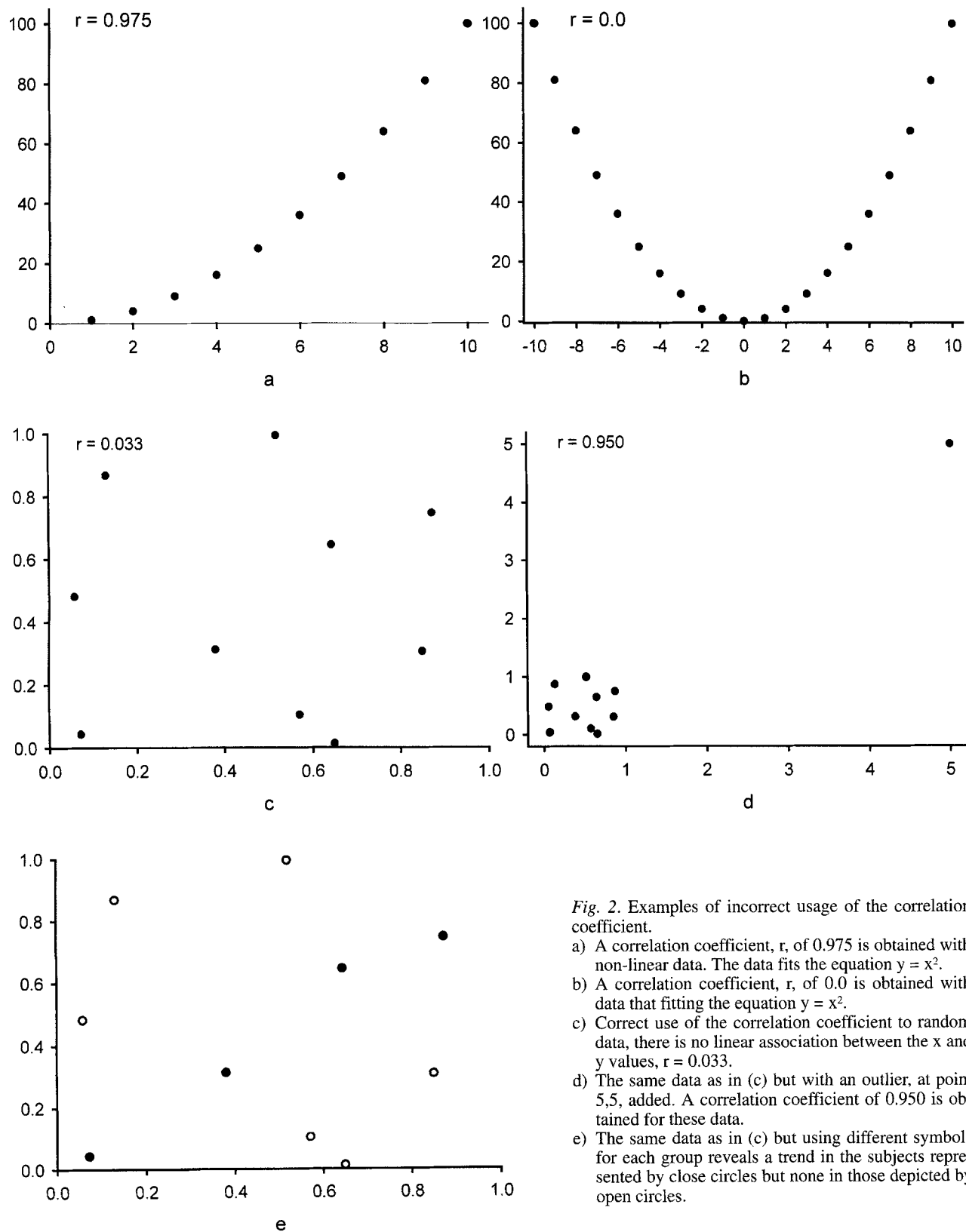


Fig. 2. Examples of incorrect usage of the correlation coefficient.

- A correlation coefficient, r , of 0.975 is obtained with non-linear data. The data fits the equation $y = x^2$.
- A correlation coefficient, r , of 0.0 is obtained with data that fitting the equation $y = x^2$.
- Correct use of the correlation coefficient to random data, there is no linear association between the x and y values, $r = 0.033$.
- The same data as in (c) but with an outlier, at point 5,5, added. A correlation coefficient of 0.950 is obtained for these data.
- The same data as in (c) but using different symbols for each group reveals a trend in the subjects represented by close circles but none in those depicted by open circles.

against an existing method. We would be very surprised if these methods were not highly correlated. A plot of the difference between method

readings against the average of the readings, known as a Bland-Altman plot should be used (10, 11).

When data are displayed and a correlation coefficient presented, it is not correct to include a regression line in the diagram as shown in Figure 3a. This is because with regression we are testing for the dependence of one variable on another and from this we may make predictions. The diagram with the regression line implies more of an association that might be the case. The correlation coefficient $r = 0.32$ with a P value of 0.047 is significant at the 5% level. The 95% Confidence Interval for this correlation coefficient ranges from 0.005 to 0.577. The amount of variation in one variable that is 'explained' by the other variable is given by the square of the correlation coefficient, r^2 . So in this example $r^2 = 0.32^2 = 0.102$, this means that only 10.2% of the variation in the data can be explained by a linear association between these two variables. Therefore, 90% of the variation in the data is unexplained! One of the problems when using correlation coefficients is that large samples can give significant results where very little of the variation of the data is explained. Therefore some consideration must be given to the size of the correlation coefficient as well; in fact, some statisticians argue it is the only statistic to consider.

There is some confusion as shown by the figures in various publications as to when to use correlation and when to use linear regression. Campbell and Machin (1) have provided some useful guidelines.

Use regression when:

1. You wish to predict one variable from another.
2. One of the variables can be predetermined or altered.
3. It is clear which is the dependent variable.

Use correlation when:

1. You want to assess the strength of an association between variables.

Generally the x and y axes are interchangeable in correlation whereas in regression they are not.

Multiple testing

One problem that can occur when analysing data is that many hypothesis tests are performed. If the significant level is set at 5% and we perform twenty *t*-tests we would expect 1 test to be significant due to random chance. The Bonferroni correction should be applied in this case (12), which essentially entails adopting a higher level of significance. If we were prepared to accept a significance level of α , and n tests were performed, we would only declare the result to be significant if the P value was less than α/n . So, if you wish to test five hypotheses (perhaps five

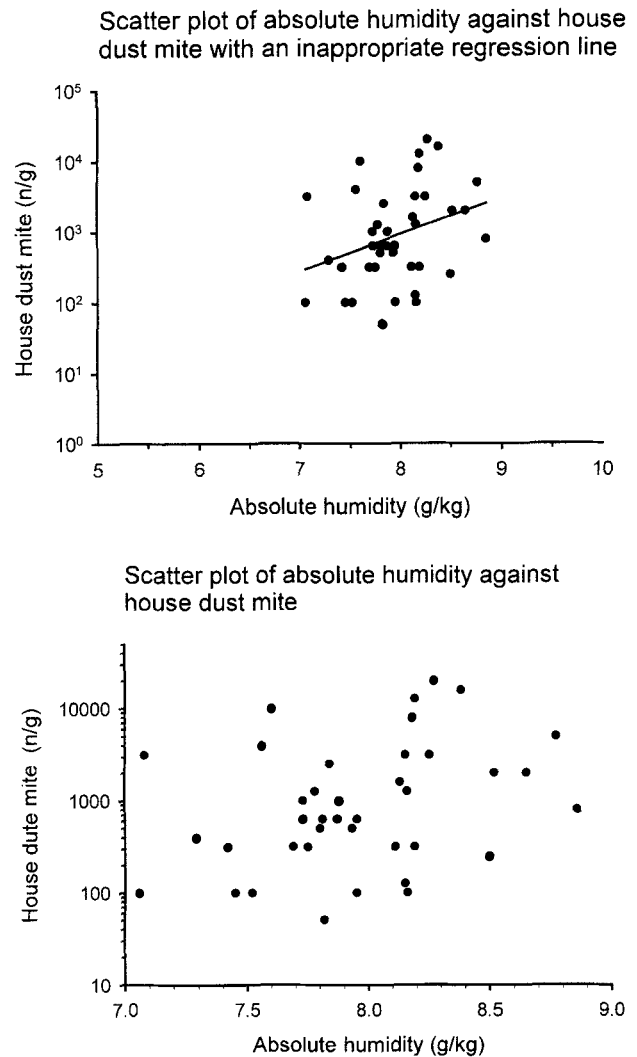


Fig. 3. a) A badly presented diagram of Absolute Humidity (g/kg) plotted against House dust mite levels (no/g) using a log10 scale. A correlation coefficient was calculated for this data and presented in the text but the statistics for the added regression line are not presented. The visual impact of the regression line together with the use of excessive white space implies a greater linear association than in Fig. 3b. b) The same data as in (a) redrawn without a regression line and using different scale so that the data points fills the plot area. This diagram implies less of an association.

different groups against a control group) then only results where $P < 0.01$ would be declared significant. It is important to note that a result where $P = 0.01$ using the Bonferroni correction is significant only at the 0.05 level. The Bonferroni method is conservative and because tests often are not independent, some real effects may not be detected (13). The method is not appropriate for analysis of repeated measures.

Repeated measures

A repeated measures experiment involves the monitoring of subjects over a period of time. Often these results are presented as shown in Figure 4. It is not unusual to find diagrams of this type with asterisks used at each time point where the authors have tested and 'detected' a significant difference. This approach is incorrect. The detailed reasons (1) can be summarised as:

1. Analyses often assume that the time points are independent, yet successive observations on a subject will be related and most analyses ignore this fact.
2. The addition of one standard error bar incorrectly implies the true curve could be drawn through any points within this range. But the probability of this is one standard error (68%) raised to the power of the number of points, in this case 7, $= 0.68^7 = 0.067$ which is small.
3. The average curve is calculated from individual curves. It displays a gradual change which could be as a result of averaging process that hides sudden shifts occurring at different time points.
4. The multiple testing may have been done to answer the question 'when does the response become significant?' which is incorrect logic. If you are monitoring continuous variables they are unlikely to suddenly change from one state to another.

Methods for analysing serial measurements in medical research have been discussed by Matthews *et al.* (14). They recommend the use of summary measures which, depending on the type of outcome measure, may involve comparison of overall means, area

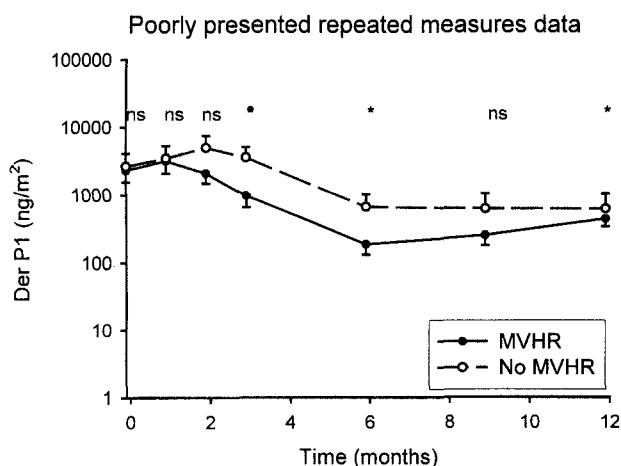


Fig. 4. A badly presented figure of repeated measures data. The inclusion of statistical significance as 'ns' and * meaning $p < 0.05$ summarises the result of testing for differences at each time point. This is an incorrect approach to analysis of this type of data. A summary measures approach should be used. Error bars should not be drawn.

under the curve, time to maximum, slope of a line, final value of outcome measure or the difference between the last and first values. They also recommend that the raw data should be plotted. A *t*-test is often appropriate in simple cases where the analysis does not require that other factors, such as age or sex, should be taken into account. Frison and Pocock (15) recommend that the use of analysis of covariance (ANCOVA) as the best technique for testing the difference between groups using summary measures. ANCOVA takes into account the between patient variations in baseline measurements when comparing post intervention measurements. This method was shown to be superior both to a simple analysis, using the mean of each subject's post intervention measurements as the summary measure, and comparison of the difference between the post treatment measurements and mean of the baseline measurements. As Campbell and Machin have said "the analysis of repeated measures is quite tricky, and a statistician should be consulted early in the process" (1).

Conclusion

Altman (16) highlighted what he termed "The scandal of poor medical research," pointing out that many bad papers have been published where inappropriate statistical analyses have been performed. As he pointed out, "it is widely considered acceptable for medical researchers to be ignorant of statistics. Many are not ashamed (and some seem to be proud) to admit that they 'don't know anything about statistics.'" This article has pointed out some of the misapplication and incorrect reporting of statistical analyses that occur in the area of pediatric allergy and immunology and we hope that in the future manuscripts submitted to the journal will not contain such errors.

Acknowledgements

I would like to thank Mike Campbell, David Machin and Mark Mullee for helpful comments on earlier drafts.

References

1. CAMPBELL MJ, MACHIN D. Medical statistics – A commonsense approach. 2 ed. Chichester: John Wiley & Sons 1993.
2. GARDNER MJ, ALTMAN DG. Statistics with confidence – confidence intervals and statistical guidelines. London: British Medical Journal 1989.
3. ALTMAN DG. Practical statistics for medical research. London: Chapman and Hall 1991.
4. CLEVELAND WS. The elements of graphing data. New Jersey: Hobart Press 1994.
5. WALLGREN A, WALLGREN B, PERSSON R, et al. Graphing statistics & data. Thousand Oaks: Sage Publications 1996.

6. INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITORS. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *Ann Intern Med* 1997; 126: 36–47.
7. BEGG C, CHO M, EASTWOOD S, et al. Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *JAMA* 1996; 276: 637–9.
8. MACHIN D, CAMPBELL MJ, FAYERS PM, PINOL APY. Sample size tables for clinical studies. 2 ed. Oxford: Blackwell Science 1997.
9. CARLSEN KH, HALVORSEN R, PETTERSEN M, CARLSEN KC. Inflammation markers and symptom activity in children with bronchial asthma. Influence of atopy and eczema. *Pediatr Allergy Immunol* 1997; 8: 112–20.
10. BLAND JM, ALTMAN DG. Statistical methods for assessing agreement between two methods of clinical measurements. *Lancet* 1986; 307–10.
11. BLAND JM, ALTMAN DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995; 346: 1085–7.
12. BLAND JM, ALTMAN DG. Multiple significance tests: the Bonferroni method. *Br Med J* 1995; 310: 170.
13. PERNEGER TV. What's wrong with Bonferroni adjustments. *Br Med J* 1998; 316: 1236–8.
14. MATTHEWS JN, ALTMAN DG, CAMPBELL MJ, ROYSTON P. Analysis of serial measurements in medical research. *Br Med J* 1990; 300: 230–5.
15. FRISON L, POCOCK SJ. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Stat Med* 1992; 11: 1685–704.
16. ALTMAN DG. The scandal of poor medical research [editorial]. *Br Med J* 1994; 308: 283–4.