Common errors in statistics and methods

Peter Flom ,[1] Katie Harron ,[2] Javier Ballesteros ,[3,4] Chester Kalinda,[5,6] Eirini Koutoumanou ,[2] Jeremy Miles ,[7] Sarah Jane Nevitt ,[8] Peter Rohloff [9]

¹Peter Flom Consulting, New York, New York, USA
²Great Ormond Street Institute of Child Health, University College London, London, UK
³University of the Basque Country, Leioa, Spain
⁴Centro de Investigación Biomédica en Red Salud Mental, Instituto de Salud Carlos III, Madrid, Spain
⁵Bill and Joyce Cummings Institute of Global Health, University of Global Health Equity, Kigali, Rwanda
⁶School of Nursing and Public Health, Discipline of Public Health Medicine, Howard College Campus, University of KwaZulu-Natal, Durban, South Africa
⁷Keck School of Medicine, University of Southern California, Los Angeles, California, USA
⁸Centre for Reviews and Dissemination, University of York, York, UK
⁹Center for Indigenous Health Research, Wuqu' Kawoq | Maya Health Alliance, Tecpán, Chimaltenango, Guatemala

Correspondence to
Dr Peter Flom; peterflomconsulting@mindspring.com

ABSTRACT

As statistical reviewers and editors for BMJ Paediatrics Open (BMJPO), we frequently see methodological and statistical errors in articles submitted to our journal. To make a list of these common errors and propose suitable corrections, and inspired by similar efforts at other leading journals, we surveyed the statistical reviewers and editors at BMJPO to collect their 'pet peeves' and examples of best practices. (1, 2) We have divided these into seven sections: graphics; statistical significance and related issues; presentation, vocabulary, textual and tabular presentation; causality; model building, regression and choice of methods; meta-analysis; and miscellaneous. Here, we present the common errors, with brief explanations. We hope that the guidance provided here will help guide authors as they prepare their submissions to the journal, leading to higher quality and more robust research reporting.

GRAPHICS

Pie charts
Pie charts are rarely the best means of data visualisation. With very few categories, it is better to use a table or just text, with more categories, a dot plot is much more effective.[1]

Dynamite plots
Dynamite plots are bar plots of numerical data, with a whisker and line showing mean values and some measure of uncertainty, often the SE. They are so called because they resemble a dynamite charge with a plunger, but, like dynamite, they often blow up. Koyama lists four big problems, which make it difficult to draw any conclusions from dynamite plots: (1) these plots are very inefficient and show too little information (2) means alone are often not that informative (3) the whiskers get in the way and can distort the overall interpretation of the plot and (4) the actual data are not shown. Better alternatives that show more of the data behind the plots, as well as their distributions, include a strip plot (ie, a plot showing each data plot, if the sample size is not very large), box plot (ideal for comparison of several groups with large sample sizes), violin plot, or bean plot.

Double axis graphs
These are graphs with two different y axes, one on the left and one on the right, for two different variables. There are two sorts of problems with these graphs: one is that they often do not clearly show what the authors are most interested in, the other is that they are easy to manipulate by slight changes of the axes. If you change the range of either axis, the appearance will change radically. It is hard to estimate the difference between the two series, or their ratio, from the graph. The suitable alternative depends on what you want to show. It might be a plot of the difference or the ratio, or two separate panels (perhaps in a lattice).

Histograms
William S. Cleveland, one of the true experts on statistical graphics, said the following about histograms:

> The histogram is a widely used graphical method that is at least a century old. But maturity and ubiquity do not guarantee the efficacy of a tool (p8).[2]

The appearance of a histogram can be strongly affected by both the starting value and the bin width. Additionally, multiple histograms do not allow easy comparisons. Better alternatives include density plots (perhaps with a smoothed line added), quantile-quantile (q–q) plots or matrices of these (to compare distributions), or quantile normal plots (to compare to a normal distribution).

Stacked bar charts
Stacked bar charts present frequencies of combinations of two categorical variables. The goal is usually to be able to compare across categories, but the fact that the various bars will not align horizontally makes this comparison difficult. Better alternatives include side-by-side bar charts, line plots (particularly if the X axis displays time) or mosaic plots.

Presenting specific results using only graphs
Two of the most common ways of presenting data and results are tables and graphs. These two methods serve different purposes and

attempts to make one meet the purposes of the other are not optimal. If you want to focus on the presentation of specific values, use tables. If you want to show general patterns and relationships, use graphs. However, tables with many rows (more than about eight) can be hard to read and might be better in an appendix. This is especially true of tables that span multiple pages.

STATISTICAL SIGNIFICANCE AND RELATED ISSUES
P values without effect sizes

Authors often present p values without the accompanying effect size, perhaps to reduce word count. With word count in mind, when choosing between an effect size or a p value, we would opt for leaving out the p value. To understand why, we must consider what p values and effect sizes are. A p value is the probability that, if the null hypothesis is true in the population from which your sample is randomly drawn from, you will get a test statistic (R or β or OR, eg) at least as extreme as the one you observed. These can be very important in some cases, as when Ronald Fisher was testing different fertilisers.[3] But they are overused and often used inappropriately to dichotomise research into 'positive' and 'negative' findings based on p value thresholds (ie, the holy grail of a p values of less than 0.05). In medicine, we often know beforehand that the null is not true and, therefore, the conditions for p values are often not exactly met. An effect size is just what it sounds like: the size of an effect. How much longer do patients live? How much is the risk of disease reduced? They are of obvious importance in pretty much all areas of research. In 2016, the American Statistical Association issued a 'statement on p values' that was along the same lines as the above and is an important reference.[4]

Effect sizes with CIs

Effect sizes should be presented with CIs to further elevate the generalisability of the results. By definition, statistical inference is imprecise. CIs provide a tangible and easily interpretable measure of precision that can be assigned to pretty much any effect size (as the current computational power has made techniques such as bootstrapping widely accessible).

Use of approximate p values

Authors may state p values in approximate terms, often using inequalities such as p<0.05. This is a holdover from the days before the easy availability of computers and statistical programmes, when you had to look up your test statistic in a table or textbook and usually could not get an exact p value. Now that you can give exact values, there is no reason to use approximate values when exact ones are available.

One exception is for very small p values, when even the computer programme may not be able to calculate p exactly. For these very small p values, it is also important to not report these as '0.0000'. A p value cannot be exactly

0 (neither exactly 1). Some software may output this due to rounding (although it should not) but what it means is that the p is less than some value, often 0.001 or 0.0001. These should be presented as, for example, '< 0.001'.

Use of the term 'insignificant'

This has the same problems as approximate p values, only worse if you are just saying p>0.05. If p values have meaning, then they have meaning across the whole range from 0 to 1. P values of 0.053 and 0.8 are both greater than 0.05, but reporting them exactly will enable the audience to draw their own conclusions regarding proximity to the relevant threshold. To save some space and repetition, results of several large insignificant p values can be reported with a bulk statement of p greater than a universally accepted large threshold level (eg, p>0.4).

P values in 'table 1'

It is common to see p values in "table 1" of a paper where basic sociodemographic and clinical characteristics of the different study population subgroups are shown, but this is rarely useful. In the case of a randomised clinical trial, taking p values is often justified to demonstrate that the randomisation 'worked'. But since you are taking p values across multiple comparator variables, something will likely turn up at <0.05 just by chance. And, importantly, it is still a randomised trial, and these values should not change your prespecified analysis plan, so just leave the p values out! In an observational trial, what matters more than the p value is the size of differences between groups, or your expert opinion on the importance of a predictor. Taking p values in table 1 leads to the temptation of letting these p values guide your analysis approach, for example, to use p value as a screen for building your regression model, which is not the correct approach.[5]

File drawer problem

A related issue is the 'file drawer problem'. This occurs when authors only submit significant findings. Although this does not affect the correctness of a particular paper, it does affect the overall literature by giving an overly strong impression of the evidence. For an extreme example, if researchers ran 20 tests of a hypothesis where the null was true, then, on average, one would be significant. If only this one was submitted and published, the effect might be regarded as backed by evidence.

PRESENTATION, VOCABULARY AND TEXTUAL AND TABULAR PRESENTATION
Multivariate versus multivariable versus multiple versus multilevel

'Multivariate' regression should be used when you have more than one dependent variable. The much more common case where you have multiple independent variables should be referred to as 'multiple' regression. 'Multivariable' could refer to either and is best avoided

completely. Multilevel regression refers to hierarchical modelling/random effects models where dependency is evident across the values of the dependent variable.[6]

Use of ±

Authors often use the shorthand notation '±' after a summary measure to represent precision or dispersion. However, the notation '±' itself is meaningless unless the actual measure referred to is explicitly mentioned, for example, SEs, SD or CIs.

Lack of detail about advanced or new statistical methods

When you use a new medical technique, for example, a new drug, or a new surgical procedure, you describe it in detail. You do not need to do this with very common techniques. The same holds for statistical methods. Consider what your audience will understand; most readers of *BMJ Paediatrics Open* are medical professionals, not statisticians. So, while you do not need to give details of how a t test is done, you should give details of how unfamiliar or obscure methods are done. It is also good to give a citation to a paper that describes the method in detail. Often, authors who are using a relatively new or uncommon statistical method will cite a paper that uses the method. It is much better to cite a paper (or *the* paper) that describes the method. Also, do not cite papers that are inaccessible, incorrect (this does happen!) or recommend *against* the use of the cited method (oops).

F statistic with no df

The F statistic is very commonly used, most frequently within the results of an ANOVA test or multiple regression. But F, by itself, means relatively little. Almost always, the F statistic will be accompanied by a p value, which is what the bulk of the readers of this journal will turn to for interpretation. If, however, you decide in favour of the reporting of F statistic, it needs the df (numerator and denominator) to be meaningful. Without that, it is like reporting a distance without specifying miles, kilometres or light years.

Unclear description of what software was used

Just as you should tell the readers which company made a tool or a drug that you used, you should tell us what statistical software was used, in enough detail that someone could then go find it. Not 'R' but 'R V.3.12'. Also, do not get the front-end mixed up with the software. For example, RStudio is an editor and R is statistical software. In addition, user-contributed or specialised add-on packages for specific statistical techniques need to be cited, in addition to the base software package used.

Unclear description of power analysis

When you do a power analysis, you should tell us what you did in enough detail that we could replicate the calculation. Typically, this requires stating some of the following: effect size, the statistical test that was used, the p value, the power desired and the sample size. You should also tell us the software that was used. If you did

power analysis by simulation (and this should be done more often than it is), you should give us the details.[7] In addition, justification of your presumed effect size should be explicit through reference to preliminary data or citations to papers that describe settings, which are representative of your own setting.

CAUSALITY

Be careful with causal language such as 'cause', 'effect' and so on. Only certain forms of research design and research methods let us attribute causality. Avoid making sweeping conclusions from observational studies. An observational study may show an association between parental education level and child malnutrition; it does not causally follow that increasing parent education will eradicate child malnutrition. We often decline to send papers out for peer review when they make this basic error.

MODEL BUILDING, REGRESSION AND CHOICE OF METHODS
Inappropriate sensitivity analyses (eg, exclusion of 'outliers', 'leave one out' analyses in meta-analysis)

'Sensitivity analysis' is used in a wide variety of ways in different fields; even within medicine, a quick Google search finds a bewildering set of examples. In general, sensitivity analyses should be conducted to investigate the robustness (Note that this is robustness in a statistical, rather than a medical sense. That is, resistance to outliers) of analysis results where assumptions may have been made within statistical methods, or in the presence of 'problems' within the analysis (eg, missing data). Like all statistical methods, sensitivity analyses should be prespecified and specific. Data-driven sensitivity analyses, such as 'leave one out' analyses in meta-analysis and exclusion of observed 'outliers' from datasets are generally not recommended as such analyses may result in exclusion of valid data and selective reporting and increase the risk of statistical type I error where multiple sensitivity analyses are conducted. Post hoc sensitivity analyses must be carefully justified and be sure to say what, exactly, you are testing.

Variable selection via stepwise, backward, forward, bivariate screening

Variable selection in regression models is part art and part science and a variety of methods can be used. But one thing is clear; the methods in this topic heading are generally not good! They result in p values that are too low, SEs that are too small and parameter estimates that are biased away from 0.[5 8] Ideally, you would use expert knowledge to select variables, but if you must use an automated method, LASSO is not bad.

Overuse of linear regression

Linear regression is one of the most common statistical methods and it has many legitimate uses. But there are many tools that were either recently invented or that recently became practical because of increases in

computer speed, where 'recent' may be 'last 50 years'. Some examples are multivariate adaptive regression splines, all sorts of regression trees and related methods, quantile regression, 'big' data (where 'big' keeps on changing), and permutation and randomisation tests. Authors should consider these advancements when deciding how to apply regression to their data.

Using inappropriate methods because 'that is what everyone does'

In the first author's consulting business (before retiring), he would often recommend an unfamiliar method. The client would listen, agree that the new method would be better and then tell him to do it the old way 'because that is what everyone does'. By this logic, we would still be following Galen and using bloodletting as a medical treatment. We see many examples of this in submitted papers. Many of these are listed in the two paragraphs immediately above, here, we want to simply emphasise the poorness of this reason. One example from the first author's experience is in neonatology, where there is a lot of interest in predicting (very) low (and sometimes high) birth weight. The usual methods are either to do Ordinary least squares (OLS) regression on birth weight or to categorise weight into two, three or four categories and do some kind of logistic regression. Here, quantile regression would be better, and the results of using it are quite different from either OLS or logistic regression.[8]

Categorising continuous variables

Categorising continuous variables increases statistical type I and type II errors and introduces a kind of 'magical thinking' that something interesting happens right at the cut points. Do not do this. For example, we edit a lot of papers on child malnutrition, where stunting is defined at <−2.0 SD. There is no substantial difference, however, between a child at −1.99 SD and another at −2.01.[9]

These statements apply to *analysis*. It may be necessary to use categories in *presentation*. In medicine, dichotomous decisions often have to be made. However, these decisions are usually based on several strands of evidence, and that evidence is best gotten from analysis that does not categorise. For example, the decision to discharge a patient from the hospital may be based on evidence from multiple tests and several doctors.

OVER-RELIANCE ON RATIO EFFECTS INSTEAD OF DIFFERENCE EFFECTS TO INTERPRET DICHOTOMOUS OUTCOMES

For dichotomous outcomes, although relative risks or ORs are most common, they are more difficult to translate in common sense, absolute terms to the population. Risk difference has poorer mathematical properties but better intelligibility. It is best, therefore, to provide both ratio and difference measures to improve data interpretation. Assuming, for instance, that the risk (probability) of improving in the experimental arm is 0.5 but only 0.3 in the control arm, the relative risk would be 1.67 meaning a 67% 'relative improvement' in the experimental arm versus the control arm. However, the risk difference of 0.2 means that 20 'more people' over 100 treated, or 20%, will improve, an absolute improvement that is more understandable than the relative risk. In meta-analytical uses, it has now become practically compulsory to use both metrics to complement the interpretation of intervention effects, and we believe it is time for this reporting standard to be applied also to all primary reports on the efficacy of interventions.

Lack of detail on how missing data are handled

Missing data are ubiquitous. Research subjects refuse to answer questions, or they drop out, or data are lost, or whatever. Do not neglect to tell readers how you dealt with this. Although the proper procedure depends on the details and should ideally be prespecified within a statistical analysis plan, one common set of procedures is multiple imputation, which is underutilised and increasingly straightforward with modern statistical software.

META-ANALYSIS

Quoting Preferred Reporting Items for Systematic Reviews and Meta-Analyses as a guideline to conduct systematic reviews and meta-analyses

In the methods section of systematic reviews, it is common to find a sentence like this one: '…this review was conducted according to PRISMA guidelines'. This sentence not only misinterprets Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) but also raises suspicions among reviewers about other methodological issues in the manuscript. PRISMA, as the name suggests, it is a guideline for reporting, not for conducting systematic reviews and meta-analyses.[10]

Misinterpreting I^2 as an absolute measure of heterogeneity in meta-analyses

The I^2 is a relative measure that indicates the variability of effects across studies as a percentage of the total variability due to statistical heterogeneity in the set of studies included in a meta-analysis and it should be interpreted as such. It is not an absolute measure of heterogeneity like τ^2 and should be interpreted more as a measure of inconsistency rather than heterogeneity.[11] A low value of I^2 does not necessarily indicate that there is no variation in the studies and is not a reason to fail to explore such variation. In addition, establishing general thresholds for which value of I^2 should be considered high or low is not recommended. The values of I^2 in a meta-analysis are influenced by the number of studies included in the analysis as well as the direction, magnitude and precision of effect sizes within those studies.[12] Therefore, while two meta-analyses may be associated with the same or similar I^2 values, the impact of clinical and statistical heterogeneity on the pooled estimates in those two meta-analyses may be very different.

Using I^2 cut-offs to choose between fixed effect and random effects

For example, a statement such as: 'if I^2 is 50% or above, heterogeneity is 'significant', and we will use random effects'. It is almost never wrong to use a random effects meta-analysis over a fixed effects meta-analysis. If there is no heterogeneity, the random effects model and fixed effects model will give the same (or very, very similar) results. If there is heterogeneity, the random effects model is the more appropriate.

Despite the above information regarding the appropriate interpretation of I^2 values, it is common practice to interpret I^2 values as absolute indicators of heterogeneity and select the model for combining effects based on thresholds of these values. The rationale behind this is that an I^2 value exceeding a certain threshold (eg, 50% or 75%) suggests significant heterogeneity among study effects, necessitating the use of a random effects model. Conversely, an I^2 value below a certain threshold (eg, 25%) indicates a consistent effect that can be appropriately analysed using a fixed or common effect model. However, the choice of model for combining studies should not rely on the observed I^2 value, but rather on the inferences to be drawn from it and the presence of clinical heterogeneity, which is usually defined by the breadth (or conversely the narrow focus) of the research question. Utilising a fixed or common effect model is suitable for estimating the effect within a specific set of homogenous studies, while a random effects model is more appropriate for estimating how the effect may vary across the population from which the studies are drawn. In conclusion, the selection of the combination model should be prespecified and based on the intended inference, rather than simply on whether the observed I^2 value surpasses, or not, a certain threshold.

Use of a fixed effects model in meta-analysis because of common population effect size

The fixed effects model is often used because it assumes that the selected studies in the meta-analysis calculate a common population effect size, and any observed difference between the studies is due to sampling error.[13] However, this is rarely the case, except perhaps for a meta-analysis of pure replications.

Quality assessment of studies

Quality assessment should be performed as a quantitative assessment of the studies included in meta-analysis using a set of predefined safeguards to reduce the risks of bias. Quality assessment tools should be used as they were designed to be used, with any user-defined adaptations clearly justified. Any issues of quality or potential risk of bias, which may impact on meta-analysis results, should be briefly described. Quality 'scores' should only be assigned according to scoring systems defined by the tool, and user-defined scores (ie, user-defined thresholds for high, moderate, low quality, etc) should be avoided. While a study with the best 'score' does not ensure high quality, it ensures that predefined safeguards to reduce risks of bias are met.

Lack of grading the certainty of the main results in meta-analysis

It is common practice to accept the combined results of a meta-analysis at face value and categorise them as significant or non-significant without considering the importance of the findings as well as potential biases and flaws in both the individual studies included in the meta-analysis and the overall estimation (particularly heterogeneity, publication bias or small study bias). To be valuable in clinical practice, the results of a meta-analysis should also be evaluated for certainty and quality using methods such as GRADE.[14]

Incomplete info to understand/reproduce the search results in meta-analysis

This is another illustration of the general rule that your paper should have enough information to allow other researchers to duplicate what you did. Some authors simply list terms (and not search engines or detailed search structure) or do not give dates.

MISCELLANEOUS
Inappropriate descriptive statistics

Just as linear regression is not the only method of regression for a continuous outcome, the mean and median are not the only measures of central tendency. There are also trimmed and winsorized means, geometric means, harmonic means and more. Sometimes, though, there is no good single numerical measure of central tendency, and you need something like a five-number or seven-number summary or a density plot.[15]

Another example is measures of spread that do not match measures of central tendency. Often, authors will recognise that the mean is not the ideal measure of central tendency for a variable. But sometimes, they give the SD anyway. The SD depends on the mean, so, if the mean is not a good measure, the SD cannot be either. Alternative measures of spread include the range and IQR, quartiles and percentiles.

Odds versus risk

Odds is not the same as risk and an OR is not the same as a risk ratio.[16]

When the outcome is relatively common, these two measures can be very different from each other, so while either one can be used in most scenarios, it is important not to confuse the two and to ensure that the reader understands the distinction. It may also be worth reporting summary statistics such as the risk difference or the number needed to treat. An example of the serious consequences of the misinterpretation of odds as risk was the 'pill scare' in the UK in 1995.[17] It was caused by medical professionals and lay people misinterpreting the meaning of an almost twofold increase in risk of venous

thrombosis, which sounds dramatic, but was an increase in risk from approximately 1 in 7000 to 2 in 7000. Because of this misinterpretation, many women stopped taking the contraceptive pill, and one consequence of this failure to understand was an increase in pregnancy rates (and pregnancy is associated with a higher risk of thrombosis) and abortion rates in the UK.

Lack of open access protocol to check reporting bias.
Most large clinical studies, and certainly all randomised clinical trials, should have a previously published study protocol, either in a peer-reviewed journal, repository or other stable online location. Systematic reviews and meta-analyses should be registered within registers such as PROSPERO (https://www.crd.york.ac.uk/prospero/) or the Open Science Framework (https://osf.io/) and for larger projects, publication of protocols may also be appropriate. Information on how to access study registration and/or protocols should be provided for all subsequent reports.

Significant in one group, but not significant in the other: interaction?
Researchers are often interested in whether relationships among variables are the same for different subsets of the data. One common way that they look at this is to analyse each subset separately and then compare p values, often basing conclusions on whether the relationship is significant in one subset, both or neither. Andrew Gelman wrote an important article *The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant,* which makes the point that you can get 'significant' and 'not significant' results even when the actual difference is tiny, or get both significant or not significant results even when it is large.[18] This question is better analysed by using interactions. To test for an interaction, include an interaction term in the model. To show an interaction, use graphics.

Coefficient (Cronbach's) alpha >0.7='reliable'
The use of 0.7 is often cited as being from Nunnally (1972) or from the second edition of the same book (Nunnally and Bernstein, 1994).[19] We suspect that most authors who cite this have not read the original texts, because what this text says is more nuanced, and it does not say that '> 0.7=good'. In addition, coefficient alpha has been described as 'riddled with problems stemming from unrealistic assumptions'.[20]

Not checking model assumptions or, at least, not reporting them
This error is very common. Most (maybe all) statistical methods make assumptions. The researcher needs to be aware of these, check them and report the results. For instance, multiple regression with ordinary least squares has several assumptions, including: (1) linear parameters, (2) correct model specification, (3) errors are independent and identically distributed and (4) errors are normally distributed.[21]

Inappropriate use of decimal places
If your total sample size is 105, you need not say that your sample was 55.238% women—you are giving accuracy to 1 in 100 000 people. Similarly, if you have measured age in years, and you state that the mean age was 34.561, you are presenting the reader with a mean age, that is accurate to (approximately) 8 hours. A rule of thumb that we like is to consider the level of accuracy that any reasonable person would consider sufficient, and then add one more figure. For example, for mean age in years, accuracy to within 1 year is almost certainly sufficient, so report 1 decimal: 34.6 years. For per cent women, 55.2%.[22]

Writing of numbers in as text or figures
Small numbers (less than 10, or multiples of 10) should be written as text. Instead of we observed 7 (xx%) children with malnutrition in our study, write, we observed seven children with malnutrition in our study. We observed 50 children with malnutrition in our study. For large numbers not in multiples of 10, you can write as, we observed 36 (xx%) children with malnutrition in our study. Furthermore, numbers at the start of sentences should be written as text. Thirty-six children with malnutrition were observed in our study. Alternatively, this can be written as: in our study, 36 (xx%) children with malnutrition were observed.

SUMMARY
Adherence to some basic principles of statistics practice and presentation would result in more robust findings and clearer articles.

X Peter Flom @peterflomstats, Javier Ballesteros @JBallesteros__ and Sarah Jane Nevitt @sjn_16

ORCID iDs
Peter Flom http://orcid.org/0000-0002-8419-7968
Katie Harron http://orcid.org/0000-0002-3418-2856
Javier Ballesteros http://orcid.org/0000-0002-6713-1916
Eirini Koutoumanou http://orcid.org/0000-0002-0731-4243
Jeremy Miles http://orcid.org/0000-0002-3229-6235
Sarah Jane Nevitt http://orcid.org/0000-0001-9988-2709
Peter Rohloff http://orcid.org/0000-0001-7274-8315

REFERENCES

1 Flom P. Peter Flom. The blog [internet] medium. 2018-. graphics for univariate data: pie is delicious but not nutritious. 2018. Available: https://medium.com/peter-flom-the-blog/graphics-for-univariate-data-pie-is-delicious-but-not-nutritious-4e9f59e00085 [Accessed 6 Jun 2024].

2 Cleveland WS. Visualizing Data. Summit: NJ Hobart Press, 1993.

3 Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd,

4 Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. *Am Stat* 2016;70:129–33.

5 Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Berlin: Springer, 2001.

6 Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health* 2013;103:39–40.

7 Arnold BF, Hogan DR, Colford JM, *et al*. Simulation methods to estimate design power: an overview for applied research. *BMC Med Res Methodol* 2011;11:94.

8 Flom P. Peter Flom. Towards data science [internet]: medium. 2018-. stopping stepwise: why stepwise selection is bad and what you should use instead. 2018. Available: https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-90818b3f52df [Accessed 6 Jun 2024].

9 Rohloff P, Flom P. Stunting: methodological considerations for improved study design and reporting. *BMJ Paediatr Open* 2023;7:e001908.

10 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.

11 Borenstein M, Higgins JP, Hedges LV, *et al*. Basics of meta-analysis: I(2) is not an absolute measure of heterogeneity. *Res Synth Methods* 2017;8:5–18.

12 Deeks JJ, Higgins JPT, Altman DG. Chapter 10: analysing data and undertaking meta-analyses. In: Higgins J, Chandler J, Cumpston M, *et al*., eds. *Cochrane Handbook for Systematic Reviews of Interventions version*. 64. London: Cochrane, 2023.

13 Lipsey MW, Wilson DB. The way in which intervention studies have "personality" and why it is important to meta-analysis. *Eval Health Prof* 2001;24:236–54.

14 Guyatt GH, Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.

15 Tukey JW. Exploratory Data Analysis. London: Pearson, 1977.

16 Higgins J, Deeks JJ. Chapter 6: choosing effect measures and computing estimates of effect. In: Higgins J, Chandler J, Cumpston M, *et al*., eds. *Cochrane Handbook for Systematic Reviews of Interventions version*. 64. London: Cochrane, 2023.

17 Bhathena RK. The 1995 pill scare and its aftermath: lessons learnt. *J Obstet Gynaecol* 1998;18:215–7.

18 Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *Am Stat* 2006;60:328–31.

19 Nunnally J, Bernstein L. Psychometric Theory. New York: McGraw-Hill, 1994.

20 McNeish D. Thanks coefficient alpha, we'll take it from here. *Psychol Methods* 2018;23:412–33.

21 Economic Theory Blog. Assumptions of classical linear regression models (CLRM). 2012. Available: https://economictheoryblog.com/2015/04/01/ols_assumptions/ [Accessed 6 Jun 2024].

22 Cole TJ. Too many digits: the presentation of numerical data. *Arch Dis Child* 2015;100:608–9.