

Considerations for the design, analysis and presentation of in vivo studies

Jonas Ranstam¹, Jonathan A. Cook²

1. Mdas AB, Rotfruktsgatan 12B, SE-27154 Ystad, Sweden, jr@mdas.se
2. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Botnar Research Centre Windmill Road, Oxford OX3 7LD, UK, jonathan.cook@csm.ox.ac.uk

Abstract

Important statistical principles and key methodological challenges in the study design, statistical analysis, and reporting of results from *in vivo* studies are considered along with the underlying statistical principles and practical suggestions. The importance of pre-specifying endpoints and analysis, the common underlying assumption of statistically independent observations, sample size considerations, and multiplicity issues are considered. Additionally, the reporting of results and graphical presentations are also discussed.

Introduction

Results of statistical tests appear in almost all published reports of *in vivo* studies. It is easy to get the impression from reading them that the sole purpose of these analyses is to produce p-values in order to be able to divide the experimental findings into either “positive” ($p < 0.05$) or “negative” ($p > 0.05$) results. Such an approach represents a very reductionist and misleading (1, 2) conception of statistical inference that corrupts the scientific process and needs to be replaced by a more adequate and nuanced understanding. P-values describe uncertainty, the magnitude of which depends on sample size, not practical importance or relevance, see Figure 1.

The main aims of a statistical analysis are to summarise the data and to make inferences relating to it. In the context of experimental science, inference typically relates to evaluating the empirical support for (or as is more common against) a hypothesis by quantifying the uncertainty relating to it. The simplest and most common statistical analyses produce a p-value (the probability of observing by chance the result or a more extreme one, if the chosen null hypothesis is true), which quantifies the strength of evidence against the null hypothesis given the data. Usually this null hypothesis is framed in terms of the absence of an effect of interest. A “positive” finding (e.g. that a specific exposure has a specific effect) is judged to be supported, i.e. statistically significant, when this p-value is low and below the threshold for statistical significance (typically the 5% level). However, the absence of a statistically significant result does not provide the negative (e.g. that the studied

exposure has no effect). Merely that the analysis conducted does not provide evidence to conclude there is an effect. Contrary to a common misconception, a statistically non-significant result is not sufficient on its own to claim a “negative” finding. To be able to do so requires consideration of the magnitude of an effect which the data supports. The degree of uncertainty (e.g. of the exposure's potential effect) in terms of an effect size and a confidence interval relating to it can readily be calculated. An interval excluding all practically important effects can then be interpreted as providing support for the absence of a meaningful finding. Presenting a “negative” finding is thus more complicated than it may appear. It should be noted quantifying magnitude of the measure of interest and the uncertainty relating to it (e.g. in a confidence interval) is useful irrespective of whether the p-value is significant and should be done routinely irrespective of the p-value. While knowing there is an effect is valuable, also quantifying how large it might be is more useful. Presenting a “negative” finding is thus more complicated than it may appear.

Accurate quantification of the inferential uncertainty requires a study design that eliminates or at least attenuates bias, a sufficient sample size, and the use of an adequate statistical analysis. The definition of the analysis unit, the number of exposure groups, the endpoint definitions, the number of analysed endpoints, and whether or not endpoints and statistical analyses were pre-specified prior to analysis of the data are important considerations regarding the trustworthiness of the reported findings. We attempt in this paper to give explanations and suggestions useful for designing, analysing and reporting statistical results from *in vivo* experiments. A

rational approach to statistical inference will form a better framework for reporting results.

Study design and statistical analysis

There is a close relationship between the study design and any intended statistical analysis, and this is particularly true for experimental studies; this is usually well recognized among researchers engaged in randomized trials. The relationship seems, however, to be less appreciated by researchers engaged in *in vivo* studies. It is possible to come away with the impression from reading reports of *in vivo* studies that as so much effort has been focused upon “getting hold of” data that what is then done with the data once it has been processed (the statistical analysis) is an afterthought. We describe and explain here a number of important methodological principles with relevance both to the study design and the statistical analysis.

Randomization and blinding

Randomization and blinding are the natural, and most desirable, features of any scientific experiment comparing alternative approaches whether they be medical or biological options. The purpose of randomization is to avoid systematic differences and thereby eliminate selection and confounding bias. It does not guarantee entirely similar groups and some random differences should be expected. Where ensuring similarity with regards to a particular characteristic, such as a factor which is known

to have a major impact on the outcome, is considered to be vital, stratified randomization (or even minimization, if there are several such factors) can be used.

Blinding is also important in both randomized trials and *in vivo* studies in preserving the benefits of randomisation. Preferably all persons that are involved in the handling or testing of the experimental animals or in the assessment of the outcome should be blinded to which group an animal belongs to in order to avoid that systematic differences post randomization generate bias.

Given the fundamental importance of how experiments are designed and performed with respect to bias reduction, it is surprising to find that such aspects are often not well described in published reports (3).

Pre-specified endpoints and analysis

Observational, or more specifically non-interventional, studies have by definition no room for manipulating the exposure of the participants, which makes such studies vulnerable for various kinds of bias, especially selection bias and confounding bias, and substantially less suitable for confirmation of pre-specified hypotheses as the most appropriate analysis is typically unclear at the outset which makes preservation of the the type I error rate problematic and therefore leaves the study more vulnerable to a spurious finding

In vivo studies and randomized clinical trials have in principle several important characteristics in common. They can for example both be designed to provide confirmatory results. A confirmatory trial is one which seeks to confirm whether a treatment effect exists. In a similar way *in vivo* studies can be designed to confirm or purely to explore relationships. Pre-specified endpoints and analyses and a protected type I error rate are typically considered a necessary requirement for a confirmatory evaluation. The study may also include other exploratory analyses which may not be held to the same standard but they should be clearly distinguished as such, and accordingly not given the same level of credibility.

While clinical randomized trials are regulated in detail by study protocols, and increasingly by statistical analysis plans and international guidelines, the planning and performance of *in vivo* studies are typically less developed. For example, a search in PubMed for the combined terms “trial” and “study protocol” results (29 November 2015) in 4192 hits as compared with 931133 for just “trial” (450 per 10^5), and a search for the combination “*in vivo*” and “study protocol” results in 70 hits as compared with 705585 for just “*in vivo*” (10 per 10^5).

Study transparency

One reason for publishing a study protocol or a statistical analysis plan prior to the start of the study is that this effectively eliminates the possibilities for gerrymandering (selecting according to the desired finding intentionally or otherwise) of endpoints

and analyses, which undermines the results of a confirmatory study. Other reasons are that the publication can help prevent unnecessary duplication of work, and that it can enable collaboration. Several journals therefore publish study protocols. Alternatively the study protocol could be provided on an institutional website. The registration of a trial in a public trials register such as clinicaltrials.gov also includes key information from the study protocol and makes the existence of the study openly known. Similar presentations of study protocols for *in vivo* studies would, even if such protocols were very brief in comparison to clinical trial protocols, be important in order to allow assessment of how the study was carried out against the study protocol.

The independence of observations

Most standard statistical tests are based on the assumption that the analysed observations are independent. Observations that originate from a related source are naturally more similar than those from another. For example, animals living in the same cage may in some sense tend to be more similar than those living in different cages. This can lead to “clustering” of outcome and if ignored can result in statistical analyses which are unreliable. Typically this relationship leads to a loss of information compared to the equivalent number of independent observations though this is not always the case. Analyses which can account for this are available though the more complex ones (e.g. generalised mixed models) can be quite demanding in terms of the number of clusters and observations within the clusters required to reliably run.

Some researchers misunderstand the word independence. Statistical independence is, however, a well-defined term. It describes how much information one observation conveys about another observation, which ranges from nothing (complete independence) to everything (one-to-one relationship). The intraclass correlation coefficient describes the degree of similarity between observations within a cluster. However, while statistically independent variables always are uncorrelated, the reverse is not true; no correlation does not necessarily imply independence.

Questions about observations' independence are usually closely related to the study design. The fundamental basis for statistical inference is the assumption that studied observations represent randomly sampled units from the population to which findings are to be generalised. It is not always obvious how an observation and population should be defined in practice (4 - 7). For example, in studies on bovine cartilage and multiple cartilage specimens from multiple cows, what is the analysis unit, the cow or the cartilage piece?

The short answer to this question is that it depends on the purpose of the analysis and ultimately the study's aim and objectives. When an analysis is based on the assumption of an independent and identically distributed random variable, as usually is the case in statistical analyses (e.g. Student's t-test), the unit of analysis is simply defined by the observations whose error distribution is used in the evaluation. The cartilage piece would naturally be the unit of analysis when measurement errors in observations from one specific cow are of interest, and the cow is the unit of analysis

when the sampling uncertainty related to the biological variation among cows is the main focus. Taking the cartilage piece as the analysis unit for evaluating effects of biological variation among cows implies an analysis of correlated observations, as some observations represent the same cow, others represent different cows with varying characteristics. An underlying assumption of an identically distributed random variable would not be appropriate in this situation (which commonly used statistical tests like the t-test or Pearson's χ^2 tests make). Neglecting to account for the correlated observations in the analysis could have two consequences: a biased estimation of the biological variance and a biased assessment of the number of degrees of freedom (i.e. incorrectly specifying the amount of information the data provides).

If inference on both the cartilage pieces and cows are planned, an analysis which recognizes this interdependence of data is needed. Conventional methods, such as repeated measurements ANOVA, can be used, but more complex alternatives (e.g. mixed models) have several advantages (8).

Sample size considerations

The sample size calculation is a key part of experiment design; it provides reassurance that the study is likely to be able to address the primary research question. Under a conventional approach, the sample size can be calculated given the statistical parameters adopted (type 1 and 2 error rates), the test to be carried out, the expected

control group level and the target difference to be detected. It also plays a role in clarify what a study was primarily designed to achieve by informing the interpretation of the findings accordingly. A common difficulty is choosing an appropriate target difference (“effect size”).

Multiplicity issues

The interpretation of the result of a statistical test depends on whether or not the test has been pre-specified and on the number of tested null hypotheses. When a test of a null hypothesis is performed at a significance level of α (usually 0.05), the risk of a false positive outcome of the test, the type I error rate, is α . If k independent null hypotheses are defined as a “family” of hypotheses and tested at a significance level of α , the probability that at least one of them will be false positive is $1 - (1 - \alpha)^k$. This is known as the familywise error rate (FWER). For example, when a family of 20 independent null hypotheses are tested at a 0.05 significance level, the FWER is 0.64. To achieve a FWER of 0.05 the significance level of each individual test must be lower than 0.05. The test-specific significance level can be corrected to compensate for this by account for the multiple tests. Under the commonly used approach, the Bonferroni method, it is defined simply as α/k . In the above example the corrected significance level is $0.05/20=0.0025$. These calculations are based on the assumption that the tests are independent and as such it is likely to be an overly conservative approach to use. With familywise multiplicity correction, the test-specific significance level depends on how the family is defined. For example, if only 5 of the 20

independent hypotheses in the above example were included in a test family, the Bonferroni corrected significance level would be 0.01 instead of 0.0025. The Bonferroni correction while often used is a somewhat controversial method as it tends towards being overly conservative (sometimes substantially so). The general methodology has been developed and more modern alternatives (11) can be recommended as they are less conservative (e.g. alternative FWER or false discovery rate approaches). However, multiplicity correction of the significance level always protects the FWER at the expense of the statistical power. To maintain the risk of a false negative outcome (the type II error) at the same level as without such correction, the sample size needs to be increased, which usually is an economical and logistical disadvantage. Randomized clinical trials are therefore often designed with a strategy for addressing multiplicity issues that avoids multiplicity correction (12), for example by restricting the confirmatory testing to a single primary endpoint that can be evaluated using one hypothesis test.

The definition of a test family needs to be consistent with the research question (9). A mistake can have serious consequences for the validity of the results. It is therefore important not only to clearly state both the research question and the test family defined for the correction of the significance level, but also the overall strategy for addressing multiplicity issues in the study, which can differ between tests as some require multiplicity correction and some may not. This is clearly described in regulatory authorities' guidelines on multiplicity issues in clinical trials (10) though

the issue is routinely ignored in published *in-vivo* studies.

As noted above the Bonferroni correction is an often used but tends towards being overly conservative. The general multiplicity methodology has been developed and more sophisticated alternatives (11) are preferable.

Deciding whether to use multiplicity correction is closely linked to the study's overall purpose and the research questions which together should drive the analysis strategy. The purpose of a confirmatory study is to test a pre-specified key hypothesis. A study protocol presenting the statistical analysis prior to any inspection of data is necessary, and if a single hypothesis test is not sufficient, steps to preserve the integrity of the finding need to be taken (e.g. by multiplicity correction of the individual test significance level). Analysis of secondary outcomes may not necessarily require the same approach but correspondingly the analysis findings will not be as compelling.

In contrast, an exploratory study has no pre-specified hypotheses and typically includes large numbers of data-generated hypothesis tests. Coherent multiplicity corrections are then not possible, and even if one was performed it would be unclear how the outcome of the tests of data dependent hypotheses could be interpreted as compared to tests of pre-specified hypotheses (13). In this scenario, the findings should be clearly identified as exploratory.

Irrespective of the purpose of the study, when multiplicity corrections are used, the

strategy for addressing multiplicity issues and the error rates used in the evaluation of the outcome should always be clearly presented to the reader and be well motivated. Whether or not a multiplicity correction has been made for multiple related analyses should be clearly stated as should whether hypotheses underlying the analyses were pre-specified or not. It is useful to explicitly describe the analyses accordingly and label them as “exploratory” and interpreted as such.

Reporting

The presentation of a study includes as an important part a description of the study design, how the above described aspects have been taken into account when developing the study design, and how they have been taken into account in the statistical analysis. The ARRIVE guidelines (14) provide an excellent basis for the preparation of a manuscript. Here we focus on the more statistical aspects.

A Statistical analysis section

Some authors only provide information on statistical methods in the figure legends. This is not sufficient to comply with the ICMJE recommendation (15) to “Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to judge its appropriateness for the study and to verify the reported results”. A better approach is to present the used statistical methods collectively in a statistical analysis section. A description of any assessment of the validity of

underlying assumptions which statistical analyses make carried out should also be included in order to reduce the uncertainty of the presented findings' empirical support.

Visual presentation of findings

Graphical presentations are often made using bar charts like the one in the left section of Figure 2 (sometime called dynamite plots). It is surprisingly common that the error bars are undefined (7), but when they are, they often indicate standard errors of the mean (SEM). In spite of being an uncertainty measure (with large sample sizes corresponding to a 68% confidence interval) this is sometimes mistakenly believed to describe data dispersion. The 95% confidence interval is a better measure of inferential uncertainty since its interpretation is independent of the sample size, and since it provides direct consistency with the results of a two-tailed hypothesis test. The standard deviation (SD) is a suitable measure to describe data dispersion for variables with Gaussian distribution. The interquartile range (IQR) or the range may be a more appropriate way to measure dispersion.

More generally this type of plot is a poor choice to present the data, because it hides more than it reveals. The number of observations and their distribution are not disclosed with the bar chart, and this information is important for the question of whether underlying assumptions are fulfilled. A better alternative, at least with small sample sizes, is the dot plot, see the middle section of Figure 2. With greater sample

sizes a box plot (the right section of Figure 2) may be a reasonable alternative as this at least gives information about the distribution of the observations. The number of observations and summary values (e.g. mean and SD or median and IQR) should still be presented in the text or a table.

Conclusions

When research studies are designed and their findings published, it is important to acknowledge the differences between confirmatory and exploratory studies. Study protocols (and statistical analysis plans), incorporating pre-specification of hypotheses, endpoints and analyses as well as the how multiplicity issues were addressed play an important role in confirmatory studies.

In contrast, test results from exploratory studies are not generally expected to be based on pre-specified endpoints and multiplicity corrections. The value of such findings are, on the other hand, more limited; the need for confirmation should be clearly acknowledged.

The value of “negative” findings depends predominantly on whether they represent the result from a confirmatory or exploratory analysis. Concluding the absence of a positive effect requires more than a non-significant p-value. Without quantification of the underlying inferential uncertainty, a negative finding is difficult to interpret as the sample size is often small which can lead to a substantial effect being easily missed.

The presentation of confidence intervals facilitates the interpretation of results, both by enabling the distinguishing of an inconclusive result (interval which including no effect and also meaningful effects) from those which do rule out a meaningful effect, and more generally by describing the degree of uncertainty (the width of the interval) with which an effect is estimated. Routinely quantifying uncertainty irrespective of the finding's statistical significance should be carried out whenever possible.

Author contributions

Study conception and design: Ranstam, Cook

Acquisition of data: N/A

Analysis and interpretation of data: N/A

Drafting of manuscript: Ranstam, Cook

Critical revision: Ranstam, Cook

Acknowledgements

None

Conflict of Interest

None

References

1. Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med* 2005;2:e124.
2. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016doi: 10.1080/00031305.2016.1154108.
3. Macleod MR, McLean AL, Kyriakopoulou A, et al. Risk of bias of in vivo research: A focus for improvement. *PloS Biol* 2015;13(10): e1002273.
4. Bryant D, Havey TC, Roberts R, Guyatt G. How Many Patients? How Many Limbs? Analysis of Patients or Limbs in the Orthopaedic Literature. *JBJS Am* 2006;88:41-45.
5. Park MS, Kim SJ, Chung CY, Choi IH, Lee SH, Lee KM. Statistical Consideration for Bilateral Cases in Orthopaedic Research. *JBJS Am* 2010;92:1732-1737.
6. Ranstam J. Repeated measurements, bilateral observations and pseudoreplicates, why does it matter. *Osteoarthritis Cartilage* 2012;20:473-475.
7. Vaux DL. Research methods: Know when your numbers are significant. *Nature* 2012;492:180-181.
8. Brown H, Prescott R 2006 (2nd ed.) *Applied Mixed models in Medicine*. Wiley.
9. O'Brien PC. The appropriateness of analysis of variance and multiple comparison procedures. *Biometrics* 1983;39:787-94.
10. EMA. Points to consider on multiplicity issues in clinical trials. European Medicines Agency, London, 19 September 2002, CPMP/EWP/908/99.
11. Levin B. Annotation: on the Holm, Simes, and Hochberg multiple test procedures. *Am J Public Health* 1996;86:628-9.
12. EMA. ICH E9 Statistical principles for clinical trials. European Medicines Agency, London, September 1998, CPMP/ICH/363/96.
13. Bender R, Lange S. Adjusting for multiple testing – When and how? *J Clin Epidemiol* 2001;54:343-349.
14. Kilkeny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) *Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal*

Research. PLoS Biol 8(6): e1000412. doi:10.1371/journal.pbio.1000412

15. International Committee of Medical Journal Editors [homepage on the Internet]. Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals [9 January 2016] Available from: <http://www.ICMJE.org>.

Appendix - Statistical glossary

Bonferroni correction - A Bonferroni correction is applied to correct the type I error when several tests have been performed. It is based on the Bonferroni inequality, which shows that if each test is done at a size of α/n , the global type I error for the n tests is lower than α .

Confounding bias - A systematic distortion of an effect of direct interest caused by the effects of other factors, called confounders, not the direct focus of concern.

Dynamite plot - A bar graph with error bars indicating observed dispersion or inferential uncertainty.

Familywise error rate (FWER) - The probability of making one or more type I errors in a family of hypotheses for which a combined measure of error is calculated.

False discovery rate - The proportion of rejected true null hypotheses of all rejected null hypotheses.

Intraquartile range (IQR) - A measure of variability based on dividing a data set into 4 quartiles by defining 3 quartiles, Q1, Q2 (also known as the median), and Q3. The IQR is from Q1 to Q3.

Mean - A measure of the central tendency of a probability distribution or a random variable characterized by that distribution.

Median - The number separating the higher half of a data sample, population, or probability distribution, from the lower half.

Mixed model - A statistical model that contains both fixed and random effects.

Null hypothesis - A particular hypothesis under test as distinct from the alternative hypotheses which are under consideration.

Pearson's χ^2 test - A chi-squared test based on observed and expected frequencies.

Repeated measurements ANOVA - The equivalent of one-way analysis of variance (ANOVA) for related, not independent, groups.

Selection bias - The error introduced when a sample is not representative of the population for which inferences are made.

Standard deviation - The most widely used measure of dispersion of a frequency distribution, the positive square root of the variance.

Standard error - The positive square root of the variance of the sampling distribution of a statistic. This includes the precision with which the statistic estimates the relevant parameter.

Statistical analysis plan - A document that contains a more technical and detailed description of a statistical analysis briefly described in a study protocol.

Study protocol - A document that describes how an experiment or an observational study will be conducted.

t-test - A test based on the *t*-distribution which, *inter alia* is the ratio of the sample mean to its estimated standard error in samples from a normal population.

Legends

Figure 1. The sampling distributions of the mean body temperature for two hypothetical samples with a sample sizes 3 and 30 from a population with a mean value and standard deviation of 36.8 and 0.4 degrees Celcius respectively. A sample mean value of 37 degrees Celcius or greater can be expected with the probability 0.12 for the sample with a size of 3 and 0.002 for the sample with size 30. A test of the null hypothesis that the sample was drawn from a population with a mean body temperature of 36.8 degrees Celcius would then, if the sample's mean value had been 37 degrees, have been statistically significant with a sample size of 30, but not with a sample size of 3.

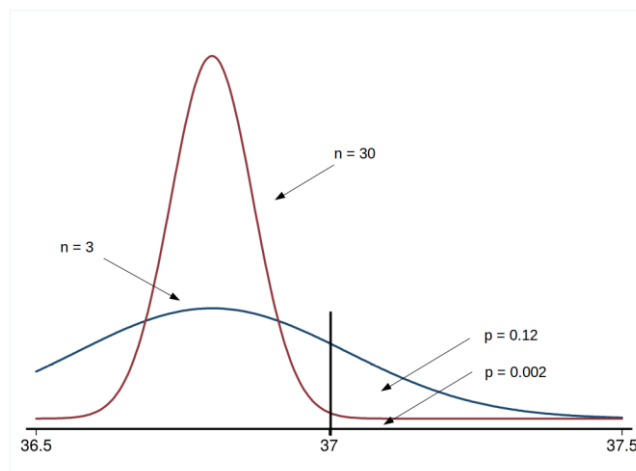


Figure 2. **Alternative graphic presentations**

