Reporting laboratory experiments

Scientific studies exhibit a growing number of advanced designs and techniques. The multiplicity and diversity complicates critical evaluation of observed data, statistical analysis, presented results and conclusions drawn.

Several reporting guidelines have therefore been developed to promote transparency in research, like Uniform Requirements for Manuscripts Submitted to Biomedical Journals (http://www.icmje.org), the CONSORT statement (http://www.consort-statement.org), STROBE (http://www.strobe-statement.org), etc. However, these reporting guidelines mainly concern clinical or epidemiological studies.

Laboratory experiments are, unfortunately, often inadequately reported. A systematic review of 44 animal studies on fluid resuscitation (1) shows, for example, that only two of the reviewed papers described how experimental units were allocated to treatment. Without this information critical assessments of the presented results' validity cannot be made.

Inadequate reporting thus devaluates the efforts and resources spent on the experiment by the researchers and on the publication of the report by editors, reviewers and readers. Osteoarthritis and Cartilage has decided to revise its information for authors and include a section with guidelines for reporting of laboratory experiments.

This is a presentation of these guidelines.

General principles

The general principle for reporting experimental studies is that the experiment should be described in a way that makes it possible for the reader to repeat it (2). The statistical analysis should also be presented with enough detail to allow a reader with access to original data to verify reported results (3).

Experimental design

Experiments should have a design that provides unbiased results with enough statistical power to detect biologically important differences or effects. The planning of an experiment should thus not only include biological or biochemical aspects; statistical methods and sample size calculations are as important.

A clear description of the chosen design is necessary for the reader's understanding of both the experiment and the statistical analysis of the data generated by the experiment.

Describe the experimental unit clearly. This is usually the smallest unit that can be independently randomized to a group, i.e. it should be possible to randomize any two experimental units to different groups. The experimental unit should also be the statistical analysis unit.

When used, describe the randomization procedure, and present the number of randomized units, replicates and number of times the experiment is repeated. If blinding is used this should also be described. If no randomization or blinding was used, state this explicitly.

Formal experimental designs, like randomized block, latin square, split-plot, etc., have been developed and are described in a number of statistical textbooks. State clearly if one of these formal designs are used. If this is not the case, describe and explained the used design in detail.

Statistical methods

Descriptions of observed data in aggregated form, for example as a mean or median value, should be presented both with a suitable measure of dispersion (e.g. standard deviation or range) and the number of experimental units included.

Observed differences or effects among experimental units should not be expected to be identical if

an experiment is repeated, because chance events always affect the outcome, at least to some degree. This creates sampling uncertainty (4), which can be quantified and should be taken into account when drawing conclusions. The experiment's results should therefore be presented with an indication of the magnitude of the uncertainty.

This can be done by considering observed data, for example a group difference or a treatment effect, as a parameter estimate and by calculating a 95% confidence interval for the uncertainty in this estimate.

Sampling uncertainty can also be evaluated using hypothesis tests. When presenting the outcome of a hypothesis test (the p-value), make sure that it is clear what hypothesis has been tested, and that the tested difference or effect (i.e. the effect size) is presented, again together with the number of experimental units included in the calculation.

Describe the statistical methods used for hypotheses testing and parameter estimation in detail. If non-standard methods have been used, give references to published descriptions, with pages stated. All statistical methods are based on certain assumptions. Student's t-test, for example, requires Gaussian distribution and homogeneous variance. If the assumptions are not fulfilled, the results may be unreliable. Assumptions should therefore be checked, and the results of the investigation should be presented.

Define also statistical terms, and specify statistical software used.

Results

It should be recognized in the results presentation that a statistically significant effect or difference is not necessarily biologically or clinically significant. It is therefore better to specify the effect size, and its uncertainty with a 95% confidence interval, than describing an effect as statistically significant, or not statistically significant.

P-values should be presented numerically, without categorization, e.g. write $p = 0.15$, not ns, and $p = 0.03$, not $p < 0.05$. When computer printout says $p = 0.0000$, write $p < 0.0001$.

Confidence intervals should be presented as (lower limit, upper limit).

The statistical power to detect an effect depends on sample size. A statistically insignificant outcome does not indicate that a tested effect does not exist; the statistical power of the test may be insufficient. Calculation and reporting of the statistical power of the experiment is thus important. The right place for presenting a-priori power assessments is in the methods section, for post hoc power assessments in the results section, and for judgements and interpretations related to power in the discussion section.

An issue related to statistical power is multiplicity. Each significance test has a chance of resulting in a false positive outcome. When more than one test is performed the overall rate of false positive tests may increase unless p-values are adjusted. Bonferroni correction is such a procedure for p-value adjustment for multiplicity.

When p-value adjustments are made, it should be clearly described what tests are included in the adjustment, and it is a good practice to present both unadjusted and adjusted p-values.

Discussion

When several independently adjusted p-values are produced, multiplicity is again created. It may therefore be useful to carefully consider a strategy for dealing with overall multiplicity and for the interpretation and judgement of p-values. If this strategy is developed prior to data collection it should be presented in the methods section. Otherwise the discussion section is the natural place for presenting the interpretations and judgements.

Potential limitations and weaknesses in study design, data collection and statistical analysis, and the

consequences of this for the validity of the findings, should of course also be discussed in the discussion section.

References

1. Roberts I, Kwan I, Evans P, Haig S. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. Br Med J 2002;324:474-476.

2. Festing MFW, Altman D. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR 2002;43:244-258.

3. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication. Available at http://www.icmje.org on June 5, 2009.

4. Ranstam J. Sampling uncertainty in medical research. Osteoarthritis Cartilage. 2009 Apr 17. [Epub ahead of print] PMID: 19410026.