Editorial

What’s in a number or in a picture?

Research findings are unavoidably uncertain. They typically rely on fragmentary information from samples evaluated using statistical inference. Statistical inference is, however, a tool for assessing the degree of uncertainty of a finding, not for proving that a finding is true. Consequently, many published findings are false (1,2).

All findings thus deserve to be treated critically, and scientific journals should try better to facilitate the readers' understanding of the findings' uncertainty (3,4). In the following we give some suggestions for manuscript authors (and readers) of *Osteoarthritis and Cartilage* in order to facilitate a better understanding of these issues.

Prompted by incidents of images in manuscripts submitted to *Osteoarthritis and Cartilage* not representing the full picture, the editors also take this opportunity to comment on current recommendations for preparing images for publication in *Osteoarthritis and Cartilage*.

Statistics in materials and methods

Statistical tests differ qualitatively from chemical tests. While chemical tests relate to the properties of a studied sample only, a statistical test concerns generalization of information collected from a sample to unstudied subjects, animals or specimens. This procedure often requires detailed information to be adequately described, much more than the name only of a statistical hypothesis test presented in a figure legend. Manuscripts should therefore, with few exceptions, have a separate statistics part in the methods section, where all statistical methods are described in sufficient detail to allow a reproduction of the results, given access to the original raw data.

When the methods used are based on certain assumptions (Student's t-test and ANOVA are for example based on the assumptions of independent observations, Gaussian distribution and homogeneity of variance), it should be investigated whether these assumptions are fulfilled or not. This investigation and its outcome should be described in the manuscript.

If serious departures from the assumptions are detected, alternative statistical methods should be considered. When observations are correlated instead of being independent, such as right and left knee of a patient, mixed or marginal models could, for example, be used. When the distribution of data is skewed, and no alternative is available, the consequences of violating the assumptions should be evaluated and described. Sensitivity tests using different scenarios is one way to do this.

Many, if not all, manuscripts include statistical analyses generating more than one p-value. Both multiple testing of the same endpoint (for example of differences between multiple groups of differently treated patients or between two groups at multiple visits during follow up) and testing of several different endpoints or outcomes is common. Repeated testing may inflate the false positive error rate substantially above the nominal significance level, and methods for correcting the p-values have therefore been developed. The Bonferroni method is probably the best known, but the methodology has been improved and better alternatives are available, such as closed testing procedures and the Sidak-Holm procedure (5).

P-values inflated by multiplicity are acceptable in exploratory or hypothesis generating studies, because the result of these studies is simply a new hypothesis, and should be declared as such. In an experiment or trial performed to confirm a pre-specified hypothesis, however, the nominal level of statistical significance is crucial; multiplicity issues must be addressed adequately.

The aim of the study (as of exploration or confirmation) should thus be made clear to the reader at the outset, and when the purpose is confirmation, a clear overall strategy for protecting against false positive errors should be presented. It will also facilitate for the reader if multiplicity corrections performed in exploratory studies are motivated.

Statistics in Results

The Materials and methods section should include only information that was available at the time when the study protocol was written or experiment was planned. Information obtained during the study should be presented in the Results section.

Variability or uncertainty - The Results section will usually include both a description of observed data and the investigators' interpretation of these data. It is important that investigators distinguish between

these two different aspects. However, many manuscripts reveal confusion instead, especially in figures. Generalizations can, for example, be presented with standard deviation (SD), which is a measure of observed dispersion, and observed data with standard error of the mean (SEM), a measure of inferential uncertainty.

Observed data should be described to show if the observations have reasonable values, if there are outliers indicating clerical errors in the data collection, if assumptions of Gaussian distribution are reasonable, if variability is homogeneous, if potential confounding factors (like age, sex, smoking habits, etc.) are balanced between groups or require adjustment of the results.

Interpretation of data, on the other hand, should take into account the uncertainties caused by sampling and measurement of the observed data, because when generalizing the results to other subjects or animals than those examined, individual variability and measurement errors will play very important roles (6). In fact, this is one of the most fundamental differences between case reports written before the 1950's and modern scientific medical reports.

The incidence of a certain disease in a group of patients or animals, the relative fracture rate among patients or animals treated or untreated with a specific substance, differences between animals of different strains, the difference in serum biomarker levels between patients with and without osteoarthritis, are examples of findings which need to be interpreted with respect to this kind of uncertainty.

The SEM is an uncertainty measure. It corresponds to a 68% confidence interval for the estimate of an unknown value. However, 95% confidence intervals are preferred, as they preserve the consistency with hypothesis tests performed at the 5% significance level, the alternative method for uncertainty evaluation.

Confidence intervals have the advantage over hypothesis tests that they describe a range of likely values. While statistically insignificant hypothesis tests only indicate of absence of evidence and not evidence of absence, the corresponding 95% confidence interval describes a range of likely values that may serve as evidence of absence of values excluded from this interval.

It should, however, be noted that graphical assessment of statistical significance between differences in group means is not unproblematic. While non-overlapping confidence intervals always corresponds to a statistically significant difference, overlapping confidence intervals do not necessarily indicate

statistical insignificance (7). The reason for the inconsistency is that distance from the mean is calculated differently in Student's t-test and in the confidence intervals. It may therefore be pertinent to present both p-values and confidence intervals, or, for avoiding p-values, to present group differences with confidence intervals for the difference in mean values, not for the two mean values themselves.

The question of whether error bars in a particular figure should reflect data description (with SD or percentiles) or interpretation (with 95% confidence intervals for mean or median values) may seem problematic to answer. However, when the purpose is to describe the characteristics of observed data, for example a Gaussian or skewed distribution, the frequency of outliers or the balance of potential confounding factors, SDs or percentiles should be used to describe variability. If the purpose instead is to describe and compare the mean treatment effects in different groups or a dose-response relation, 95% confidence intervals should be used to present the uncertainty in the estimated effects.

Significance - An analogous problem is the common and unspecific use of the word "significant", which has at least two fundamentally different meanings: a) the clinical or practical significance of an effect or difference for the subjects affected by it, and b) the statistical significance or the uncertainty of whether an observed effect or difference is caused by chance or not. A clinically significant effect may well be statistically insignificant and vice versa. Just stating that an effect is "significant" is thus ambiguous. Specify if statistical or clinical/practical significance is referred to. If clinical or practical significance is what is meant, consider using ‘relevant’, ‘important’ or ‘meaningful’ instead, and when statistical significance is meant, consider using ‘likely’ or 'probable' instead.

Results are too often presented in terms of whether a difference or effect is statistically significant or not. This leads to an artificial dichotomy of results suffering from a by multiplicity inflated false positive error rate and an ignored false negative error rate.

The general rule should be to present quantified results, estimates of true effects or differences in practical terms, e.g. length, height, width, weight, volume, etc., together with an indicator of the uncertainty of the estimate (preferably a 95% confidence interval as this describes the statistical precision better than a p-value). However, if a hypothesis test is preferred instead of a confidence interval, the p-value should be presented precisely, i.e. unless $p<0.001$ as $p=0.035$, not as $p<0.035$ or $p<0.05$ and never as $p>0.05$, n.s. or as $p=0.000$.

Number of observations - The number of observations should be clearly stated for each test or estimate. When repeated or multiple measurements are made on the same subject or animal or Petri dish, both the number of independent observations and repeated observations should be presented.

Graphs and tables – Figures are valuable when they present complex results in a readable way. However, much information can be presented more clearly and with less space in a table. The information usually described in bar charts is, for example, often better presented in a table.

The principles for presenting observations and interpretations of data, discussed above, remain, however. They are independent of whether a graph or a table is used for the presentation. We therefore emphasize: For figures the current use of S.E.M. in error bars should be replaced by 95% confidence intervals to describe uncertainty.

The picture, the whole picture and nothing but the picture?

Good science requires reliable data, and even then we may as scientists too often be proven wrong (1, 2). Images of gels, immunochemistry, microscopy, radiographs, magnetic resonance, ultrasound and more form an important part of the data presentation in many manuscripts submitted to *Osteoarthritis and Cartilage*. Prompted by incidents of images in manuscripts submitted to the journal not representing the picture, the whole picture and nothing but the picture, we take this opportunity to comment on what is commonly termed image manipulation. This appears to be disturbingly frequent, and has led to numerous editorial comments in other journals, see for example (8-12). Some instances of image manipulation may be caused by a lack of understanding or knowledge of the do's and don't's when preparing images for publication. However, this excuse is wearing increasingly thin in the current reality of widely published high profile cases of image manipulation and blot doctoring, and the immediate internet availability of author guidance in these matters, such as for *Osteoarthritis and Cartilage* http://www.elsevier.com/wps/find/journaldescription.cws_home/623055/authorinstructions .

It is simple to modify digital image files; we all have the software in our computers. But many of the modifications possible are incompatible with good science, and may represent data falsification or fabrication, leading to a serious accusation of scientific misconduct. The result might be the retraction of published work, and worse in the form of a career disruption. Co-authors share full responsibility in these cases. As noted, the editors of OAC have come across some recent instances, taking far too much time, effort and grief for all involved to resolve. Time better spent doing good science.

A good summary rule is provided by the following quote from the article by Rossner and Yamada in the Journal of Cell Biology (8): *"No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. The grouping of images from different parts of the same gel, or from*

different gels, fields, or exposures must be made explicit by the arrangement of the figure (e.g., using dividing lines) and in the text of the figure legend. Adjustments of brightness, contrast, or color balance are acceptable if they are applied to the whole image and as long as they do not obscure or eliminate any information present in the original. Nonlinear adjustments (e.g., changes to gamma settings) must be disclosed in the figure legend."

Authors are advised to read this article and the other articles cited below in full, as well as the OAC guide for authors before submission of a manuscript containing images.

And don’t forget: It is absolutely critical that you archive and time stamp your original raw data and images as they are collected. This allows you to verify accuracy should the need ever arise. Without the original data or images you are lost; should your published results ever be questioned, you cannot prove yourself right.

Jonas Ranstam

L Stefan Lohmander

Department of Orthopedics, Clinical Sciences Lund, Lund University, Sweden

Acknowledgements

Stefan Lohmander is supported by the Swedish Research Council and Lund University. The sponsors had no role in the writing or decision to publish.

Author Contributions

Both authors contributed to writing and editing of the manuscript, and approved the final submitted manuscript.

Competing Interest

The authors are deputy editor for statistics (JR) and editor-in-chief (LSL) for Osteoarthritis and Cartilage, and declare no conflict of interest.

References

1. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124. Epub 2005 Aug 30.
2. Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. False-positive results in cancer epidemiology: a plea for epistemological modesty. J Natl Cancer Inst 2008;100:988-95.
3. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol 2007;60:324-9.
4. Altman DG. The scandal of poor medical research. Br Med J 1994;308:283-4.
5. Bauer P. Multiple testing in clinical trials. Stat Med 2006;10: 871-90.
6. Ranstam J. Sampling uncertainty in medical research. Osteoarthritis Cartilage 2009;17:1416-9.
7. Julious SA. Using confidence intervals around individual means to assess statistical significance between two means. Pharmaceut Statist 2004;3:217–222.
8. Rossner M, Yamada KM. What's in a picture? The temptation of image manipulation. J Cell Biol 2004;166:11-15.
9. Neill US. Stop misbehaving! J Clin Invest 2006;116:1740-1741.
10. Neill US, Turka LA. Navigating through the gray (and CMYK) areas of figure manipulation: rules at the JCI. J Clin Invest 2007;117:2736.
11. Neill US. All data are not created equal. J Clin Invest 2009;119:424.
12. Science journals crack down on image manipulation. Published online 9 October 2009 Nature doi:10.1038/news.2009.991.