Editorial

Guidelines for a structured manuscript: Statistical methods and reporting in biomedical research journals

Robin CHRISTENSEN [1,2], Jonas RANSTAM [3], Søren OVERGAARD [4,5], and Philippe WAGNER [3,6]

[1] Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark; [2] Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Denmark; [3] Clinical Sciences, Lund University, Lund, Sweden; [4] Department of Orthopaedic Surgery and Traumatology, Copenhagen University Hospital, Bispebjerg, Denmark; [5] Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [6] Centre for Clinical Research, Uppsala University, Västerås, Sweden.
Correspondence: robin.christensen@regionh.dk

General information

These *Acta Orthopaedica* journal guidelines have been written to structure manuscripts before considering submission for the benefit of sound scientific work and to help authors prepare their manuscripts in accordance with good statistical standards. Moreover, it is anticipated that the guidelines will make the articles easier to read for the end-user. Authors submitting papers to peer-reviewed medical journals should preferably plan their statistical analyses while planning their study as part of the protocol stage. Authors should analyze data and describe their statistical methods in the final reporting with enough detail to enable knowledgeable readers with access to the original data to make valid judgments and verify the reported results (https://www.icmje.org/).

Depending on the study design, statistical requirements will differ. Randomized controlled trials typically include a certain number of patients based on sample size and statistical power considerations, whereas observational studies might need to reveal only what authors felt was a sufficient study population during the planning phase, depending on the study's context. In exploratory experimental studies, the number of units studied may be based on other considerations and still be justified. Regardless of the study type, authors should ensure that their data is of high quality, and all data should be stored securely and be retrievable upon request. Above all, the use of a statistical method presupposes appropriate knowledge and understanding so authors can produce valid results.

Reporting your study

Presentation and dissemination of results should focus on their scientific/clinical—not statistical—importance, with utmost attention to accuracy. The Declaration of Helsinki states, "*Authors have a duty to make publicly available the results of their research on human subjects and are accountable for the completeness and accuracy of their reports*." Protocols should be considered mandatory before initiating any research project; to help enforce this principle, prespecified protocols should also be available upon request. Planning the study properly requires that authors should have already consulted reporting guidelines when planning a research project (https://www.equator-network.org/). Various reporting guidelines have been developed for different study designs. Famous examples of these guidelines include the STROBE and RECORD statement for observational studies (cross-sectional, case-control, and cohort studies); STARD statement for studies of diagnostic test accuracy studies; PRISMA statement for systematic reviews and meta-analyses; CONSORT statement for randomized trials, while the SPIRIT statement relates to randomized trial protocols. When authors submit a manuscript, they are expected to report according to the relevant guideline depending on the research design (https://www.equator-network.org/).

Prevent misconduct in reporting

Independent of research design, there are several important principles that apply when reporting results in a respected journal. Unfortunately, some authors choose which data to report and how to report it after too many analyses have been performed. This choice will cause such "creatively minded" authors to deselect the outcomes and analyses that would not fit into what they feel constitutes a "significant manuscript." Data-dredging or fishing is a research practice associated with misconduct that involves manipulating or analyzing data in multiple ways until a desired or statistically significant result

DOI 10.2340/17453674.2023.11656

OVERVIEW: A structured manuscript according to generic guidance from the EQUATOR Network [a]

Title
- Identify value to the reader (e.g., be explicit about patients/population, interventions/exposures, and if possible, the major outcome of interest).
- Please add the research design to the latter part of the title (e.g., a randomized trial, a prospective cohort study, etc.).

Abstract
- Use structured summary form (Background, Purpose, Methods, Results, and Conclusion).
- Remember to add trial registration if the study was appropriately pre-registered.

Introduction
- Introduce in short, the topic.
- Scientific background for the topic (incl. gaps in knowledge).
- Evidence-based research (i.e., what is already known on this topic?).
- Rationale for this study (i.e., what are the challenges to be addressed?).
- Hypothesis when appropriate, aim and/or key objectives.

Methods section (include only what is available when planning)
- Structured reporting according to study design like STROBE and CONSORT (i.e., see EQUATOR network guidance).
- Study design.
- Participant/patient, samples (i.e., eligibility criteria)
- Interventions/exposures (i.e., describe groups of importance for statistical testing).
- Variables and outcome measures (e.g., the primary and key secondary endpoints).
- Sample size and power considerations (i.e., informative even in a retrospective study).
- Patient and public involvement in the research (i.e., did the researchers involve patients as research partners at any/all stages?).
- Ethics and study registration like Clinicaltrials.gov (i.e., ethics approval obtained and availability of pre-registered protocol).
- Statistical methods (e.g., main analyses; handling of missing data and multiplicity issues).

Results
- Participant flow (i.e., a Figure: a diagram illustrating study flow and attrition).
- Baseline characteristics (i.e., a Table format reporting descriptive statistics for all participants in the intention-to-infer from population).
- Main findings illustrated (i.e., illustration of the primary findings based on the prespecified objectives rather than chance findings [i.e., not based on significant "P values"]).
- Main analyses on the primary and key secondary objectives (i.e., Table(s) reporting statistical measures for each group and difference between them [with 95% confidence intervals]).

Discussion
- Statement of principal findings based on the aim/key objectives/hypothesis.
- Putting the research into context (to previous studies).
- Possible explanation of the results.
- Strengths and limitations of the study.
- Conclusions are strictly related to the aim/key objectives/hypotheses.
- Perspectives of the study. Avoid recommendation unless the manuscript is a recommendation paper.

[a] "Enhancing the QUAlity and Transparency Of health Research" (EQUATOR) Network is an international initiative that seeks to improve the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines (www.equator-network.org).

is obtained. P value hacking is a specific form of this malpractice, which involves adjusting or manipulating the significance level (P value) of a statistical test to reach a desired conclusion or to make a study's results appear more significant than they are. Consequently, these "creative authors" are not simply reporting facts, as proper science demands. Authors who *post hoc* modify their research objectives after seeing their results (following too many statistical maneuvers) represent a misconduct practice that is associated with scientific fraud (https://www.icmje.org/).

Reporting the Introduction

The scientific background of the study provides important context for readers. Therefore, the Introduction section should provide a short description of what is already known on the topic, the issue to be addressed and what gaps in current knowledge it would investigate. The search for prior studies' approach should be systematic. *Acta Orthopaedica* recommends that authors apply an evidence-based research (EBR) approach that includes a systematic and transparent format to explicitly include earlier studies (e.g., by referring to published systematic reviews). This means that authors should refer to prior papers (or a systematic review on the topic), and not selectively leave out individual papers because of apparently conflicting results.

At the end of the introduction section, after providing context and background for the study (that is, the nature of the problem and its clinical significance), the authors must clearly state the aim of the study and if any of the objectives will be subject to statistical testing. They should also elaborate on the rationale for the original research purpose, as well as if possible present a (falsifiable) hypothesis. Proper scientific procedure as well as ethical standards demand they do so before seeing the results of the study.

Reporting the Methods section

A well-structured methods section includes only information that was available at the time the plan or protocol for the study was being written; all information obtained during the study should be reported in the Results section. Authors should state the planned number of participants for the study and why this number was chosen. They must then describe how the patients were to be selected and the eligibility criteria that were employed in their selection. In the Results section (see below), the authors should present information on individuals/patients who declined to participate, withdrawals from the study, and participants with incomplete follow-up. Authors should describe in detail how measurements were made, and the techniques used. In the Results section, inserting flow diagrams for the above steps is recommended, as such diagrams elegantly present the progress through the phases of a longitudinal observational study. Flow diagrams are strongly recom-

mended because they are extremely effective for indicating missing data in epidemiological and clinical research. Randomized trials are not the only type of study to benefit from a transparent format that clearly describes enrollment, intervention/exposure allocation, follow-up, and data analysis.

Reporting the Statistics section

We encourage authors to recognize the importance of missing data—to embrace this issue and discuss (as part of the Results and Discussion section) how missing data affect the clinical findings. Missing data is unavoidable, but its potential to undermine the validity of research results is frequently ignored in the medical literature.

The links between the research question and its answer need to be developed prior to the statistical analysis in the form of a study design, accounted for in the statistical analysis, and explained to the reader of the submitted manuscript. The protocol, developed according to the principles stated in the Helsinki Declaration, should be registered in a clinical studies database such as Clinicaltrials.gov or EU Clinical Trials Register. Alternatively, the original protocol could be registered and made publicly available at the Open Science Framework (https://osf.io/). Registration of register studies is also highly recommended – identifying and documenting the study in a public registry before the study is conducted. Prospective registration of register studies has several benefits, including: (i) reducing publication bias by making it more difficult for researchers to selectively report only the results that support their hypothesis; (ii) improving research quality by encouraging researchers to carefully plan their study design, analysis, and reporting, which can improve the overall quality of research; and (iii) increasing trust in the register research and facilitating replication.

All statistical methods should be clearly specified and—when unusual methods are necessary—referenced. For every statistical result, the method used for deriving it should be clearly described. It is also important to address in sufficient detail the assumptions underlying the statistical methods used. No data should be removed, imputed, weighted, adjusted, or trimmed without clearly describing and justifying why and explaining the subsequent effects (i.e., see sensitivity analyses).

Descriptive statistics: Descriptive statistics form an indispensable part of medical research manuscripts. Suitable tables should clearly describe the important features of the collected outcome variables and of the key prognostic and demographic variables. The results of the main analyses relating to the objectives of the study should be clearly described and presented, with descriptive statistics detailing both the central tendency and measures of dispersion (spread) of the data. We use means and standard deviations, or medians and interquartile ranges, as well as counts and proportions to inform the reader regarding the distribution of observations in variables for analysis and reporting.

Statistical tests: The relation between the studied hypothesis and the presented results from null hypothesis testing (P values) should be clearly explained in the manuscript. The tests should be used with a defined effect size (e.g., estimating treatment effects), and the estimation uncertainty (usually via a confidence interval) should be considered in the results presentation. Unless the use of 1-sided tests is specifically justified (and performed at half the alpha level), the tests should be 2-sided. Authors should present P values with real numbers if these are greater than 0.001, using one digit except zeros. Otherwise, they should use "$P < 0.001$". Authors should not use "ns," "$P > 0.05$," or asterisks. We recommend that authors present analysis results with 95% confidence intervals instead of P values. Authors who wish to publish a manuscript with statistical tests must comply with 2 *Acta Orthopaedica* principles for concluding whether scientifically important differences exist:

1. A statistically non-significant test is not sufficient to claim "no difference." To show "no difference," a smallest clinically relevant size of the difference (it might be 0) must be defined. If all clinically relevant differences are excluded from the difference's confidence interval, a "no difference" or similarity/comparability conclusion is reasonable.
2. A statistically significant test does not necessarily imply a clinically important difference. The importance of the tested null hypothesis depends on the smallest clinically relevant difference that should be defined *a priori*. If the difference's confidence interval excludes all clinically irrelevant differences, a conclusion concerning the existence of a clinically important difference is reasonable.

Multiple statistical tests: Most manuscripts include and rely on more than 1 set of 95% confidence intervals and P values. However, performing multiple statistical significance tests increase the chance of false-positive test results. When a single statistical test is performed at a 5% significance level, there is just a 5% chance of a false-positive result, but if repeated tests are performed, each at a 5% significance level, a false-positive test result can be expected. Problems related to this inflation of the significance level are known as multiplicity issues, which need to be acknowledged in the interpretation of the research findings.

In contrast to hypothesis-generating studies, in which the outcome is a hypothesis, confirmatory studies—designed to provide empirical evidence for a prespecified hypothesis at a specific significance level—need to be designed and analyzed with respect to multiplicity issues—matters requiring multiple testing. Such multiple testing might be *"…due to multiple subgroup comparisons, comparisons across multiple treatment arms, analysis of multiple outcomes, and multiple analyses of the same outcome at different times."* The development of a prespecified strategy for addressing multiplicity issues is usually required. Such strategies are often, but not always, based on adjusting P values or significance levels using the Bonferroni method or more refined alternatives such as Bonferroni-Holm's or Hochberg's method. How-

ever, merely performing a *post hoc* Bonferroni adjustment in a hypothesis-generating study is not sufficient for drawing confirmative conclusions.

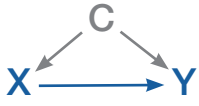
Although authors are at liberty to choose suitable significance levels, if they deviate from the conventional 5%, they should clarify their motivation and explain whether their ambition is to publish a hypothesis-generating or a confirmative study. In the latter case, and if multiplicity issues exist, they should present the multiplicity strategy they used in the Methods section and provide documentation for its pre-specification.

Multiple regression models: Authors should keep in mind when conducting regression analyses and reporting results that such analyses are conducted in different ways, with different aims in mind, depending on the design of the study. The following examples from 3 common study designs illustrate how these analyses may differ.

In prognostic studies, where the purpose is to assign future outcomes to individual patients, selection of variables to include in the model is often data-driven meaning that there is no *a priori* set of hypotheses to justify the choice of covariates with the final set of predictor variables agnostically being determined by a computer algorithm. The focus of such analyses is to optimize model performance in terms of discriminatory accuracy and/or calibration, which are often evaluated by measures such as explained variance, concordance indices, and/or calibration intercept and slope—both in the study as well as in external cohorts. The latter is key when introducing the model into clinical practice.

In contrast, intervention studies, where the purpose is instead to answer causal questions (akin to asking what happens if we treat patients with strategy A or B) have 2 main "settings:" observational and randomized. In observational studies, the main purpose of regression analysis is to reduce confounding bias in the intervention-effect estimate. As such, variable selection is primarily based on literature and/or expert opinion, sometimes encoded using the Directed Acyclic Graphs (DAGs) framework. Here, one identifies causal paths between variables that help define the confounding variables as those causing both exposure and outcome. There are also other, more pragmatic approaches to variable selection, depending on the author's level of causal knowledge. However, most data-driven algorithms mentioned in connection with prognostic studies have no place in this study setting.

The editors of *Acta Orthopaedica* recommend the following illustration of and pragmatic definitional elements of what makes a confounding variable (C):



- The Covariate (C) is an ancestor (cause) of the outcome (Y)
- The Covariate (C) probably causes the exposure (X; e.g., group or exposure)
- The Covariate (C) is not a descendant (effect) of the exposure (X) or outcome (Y)

In the randomized setting of intervention studies, which enables estimation of average causal effects without confounding bias, the purpose of regression analysis is instead to increase both the statistical precision of the intervention effect estimate and the subsequent power to detect differences between study arms. Here, one typically tries to identify from literature those variables known to be strong predictors of outcome and to predefine them for inclusion in the regression model in the study protocol.

Once the choice of variables is settled, authors should confirm that the resulting regression model accurately describes the data—that it fits. If it does not fit, authors are at risk of inducing bias in study parameter estimates, as opposed to removing it, and ultimately of reaching the wrong conclusions. The process of ensuring that statistical models fit the study data is referred to as validation and varies in nature among different regression models. Nevertheless, validation is equally important for all regression models. Given that the approach to regression analysis is different depending on the study design, authors should do the following when reporting their results:

- declare the purpose of the regression analysis;
- define what criteria are used for including variables in the regression model and how these criteria relate to the purpose of the model; and
- describe what was done to validate the model and how the validation outcome affects the interpretation of the study results. For instance, prediction models that have not been externally validated should not be recommended for clinical use.

Handling of missing data and sensitivity analyses: As stated above, missing data is unavoidable in epidemiological and clinical research and must be explained otherwise it could undermine the credibility and validity of the research results. Missing values, for either predictors or outcomes, occur in all types of medical research. Unless prompted to do otherwise, most statistical packages explicitly exclude individuals with any missing value on any of the data analyzed. The resulting so-called "available case" or "complete case" analysis is the most common "default approach" to handle missing data, although it is rarely justified. It is important that authors explicitly report how missing data was handled. As different statistical methods used to handle missing data can lead to differing conclusions, we recommend that as a minimum, sensitivity analyses be conducted to assess the robustness of the primary results.

The only kind of missingness that can be ignored (and thus excused) are the rare cases where data are *"Missing completely at random"* (MCAR)—when there are no systematic differences between the missing values and the observed values. Unlike the data that are MCAR, data *"Missing at random"* (MAR) are frequently considered the most obvi-

ous default assumption. For data that are MAR, any systematic difference between the missing values and the observed values can be explained by the observed data. For these MAR data, multiple imputation techniques are advocated as the preferred imputation method, as this approach also leads to more correct standard errors, P values, and confidence intervals. Multiple imputation essentially means creating multiple copies of the data set, with the missing values replaced by imputed values drawn from their predicted distribution by using the observed data. These multiple imputed data sets are then all analyzed by using standard procedures for complete data, followed by a meta-like technique combining the inference from these analyses. Single imputation techniques (e.g., baseline observation, or last observation carried forward) will erroneously increase the precision (too narrow confidence intervals) but these can still be highly informative as part of the sensitivity analyses.

Competing risks; Many studies focus on describing the frequency of disease, or some other negative state, in a population over time using the concept of "risk." Whereas this is a simple concept to understand and is defined as the proportion of initially disease-free individuals who develop disease (or experience an event) over the time-period, the calculations are often complicated by the fact that a non-negligible proportion of participants are lost to follow-up. For instance, participants might move out of the country before the study ends or drop out and not show up for planned study visits. Consequently, their intended time under study is only partially observed (i.e., their observations become excluded from the study and are effectively censored). Subsequently, as the simple definition of risk given above is not compatible with censoring, risks need to be estimated using more complicated techniques, such as the well-known Kaplan–Meier (K–M) method.

In some contexts, however, calculations need further considerations. One such context that has been the focal point of much scientific debate is when there is a "competing event"—when censoring is the result of events preceding the event of interest. For example, when a patient dies before their implant fails in a durability study and consequently prevents the observation of the time to implant failure for that patient, then patient death would constitute such a preceding event. The presence of competing events changes the interpretation of risk estimates and may induce bias if risks of failure differ between patients who were censored and those who were not. Alternative techniques that account for competing events may then be preferable, including basing analyses on the cumulative incidence function, and replacing the common Cox proportional hazards model with the proportional hazards model of Fine and Gray. Whereas this option is probably chosen most often, other approaches—such as using multi-state models—can also be used to analyze competing risk problems.

Although alternative approaches sometimes can be preferable to the K–M method, in several cases publications have misinterpreted the differences in results gained from the K–M method and those gained from competing risk models. Because the presence of competing events changes the interpretation of risk estimates, we advocate that choosing a risk model should be guided by the aim of the study. For instance, net failure estimated by the K-M estimator is the relevant measure when comparing the failure rates of different implants, as it would clearly not be reasonable to include effects of patient survival in this comparison. On the other hand, crude failure, estimated by the cumulative incidence method for competing risks, is the relevant measure if patient survival is part of the problem, as would be the case when studying health economics and planning resources.

When applied correctly, both estimators can be useful, and both provide unbiased estimates in the absence of confounding and selection effects. The main difference between them is that net failure refers to a hypothetical existence in which competing risks are assumed to be eliminated. In real life, the (crude) failure rate is lower for elderly patients as they are more likely to be excluded from failure because of the competing risk of dying. Whatever the choice, the competing risk problem should be acknowledged when present. If K–M estimates are presented instead of cumulative incidence estimates, the number and type of censored observations should at least be described.

Analyzing repeated measurements: Repeated measurements on the same participant are correlated and not statistically independent, so a statistical method allowing correlated observations should be used (e.g., when analyzing repeated measurements using mixed-effects models). A possible alternative would be to summarize all values from each participant into an individual estimate of a clinically relevant entity (e.g., the magnitude of a peak value, area under a curve, doubling time, etc.) and then use these estimates as input in an analysis with only one observation per participant. Again, when multiple null hypotheses are tested with the aim of confirming a prespecified hypothesis, care should be taken to avoid spurious significance by using techniques for simultaneous inference. Pre-specification is, however, necessary for confirmation. Again, the use of techniques for simultaneous inference without a prespecified null hypothesis should be explained and have a clear, valid purpose.

Using mixed models when repeated measurements are available for individual participants can also help when handling incomplete outcome data (i.e., missing data). For example, missing data can be caused by patients missing some visits or dropping out of a study. Mixed models assume that the missingness is independent of unobserved measurements but may be dependent on observed measurements. Because this frequent "default assumption" corresponds to "Missing at random" (MAR), analyzing repeated measurements using mixed models is considered comparable in validity to the multiple imputation approach (see handling of missing data and sensitivity analyses).

Reporting the Results

Authors should give numeric results not only as relative measures (e.g., percentages) but also as the absolute numbers from which the derivatives were calculated. However, if absolute numbers are given in tables, it is unnecessary to also present them in the text. Results should summarize the findings in logical sequence in the text while referring to the estimates reported in tables, and figures, with the main or most important findings presented first (preferably based on the predefined purpose). As also stated previously, authors are strongly encouraged to use a flow diagram when reporting results for all research designs (i.e., not just for randomized trials).

It is important to be clear about the order of tests; authors should infer from the primary objective (according to the original protocol) before introducing secondary findings. As a rule of thumb, only the primary objective(s) and a very few key secondary objectives are pivotal to a research paper, so findings from prespecified analyses should gain preferable mention in the Abstract.

Robustness and sensitivity analyses: Because bias can occur in subtle ways and its effects are not directly measurable, it is important to evaluate the robustness of the results and primary conclusions of the trial by running sensitivity analyses. When authors upload supplementary materials (e.g., appendix files) with several sensitivity analyses, the robustness of the findings for the primary analyses can be evaluated with greater credibility. Reassuring sensitivity analyses imply that the primary findings of the study (i.e., those explicitly reported in the manuscript) are not substantially affected when analyses are carried out based on alternative assumptions or analytic approaches. Interpretations of how treatment effect and treatment comparisons might influence statistical measures of uncertainty should also consider how bias might affect the confidence interval (and possibly P values) and statistical inference in general.

Missing data frequently represents a potential source of bias in clinical research, and it is often treated with simple (single) imputation methods involving the filling in of a single value for each missing value by methods such as the last observation carried forward (LOCF) and the baseline observation carried forward (BOCF). Such single-imputation methods should not be used as the primary analysis approach for treating missing data, but they can be informative as a sensitivity analysis.

A sensitivity analysis consists of several steps: (i) drawing conclusions under working assumptions regarding missing data; (ii) identifying a set of plausible alternative assumptions; and (iii) studying the variation in the statistical output and conclusions under these alternative settings. No matter what approach is taken for the primary analyses, we encourage authors submitting manuscripts to interpret their findings collectively supported by a series of sensitivity analyses.

The Discussion

Regarding scientific articles, *Acta Orthopaedica* would like to publish well designed, conducted, and reported studies. The wording that authors use must always strive for clarity. Please exclude unnecessarily complicated language (i.e., jargon) to impress rather than to inform the audience; authors should also be careful to avoid slang and nontechnical use of technical terms. Moreover, although technical terms might be necessary in biomedical research, authors need to invest time to make sure that the manuscript language is as clear as possible before submitting it to a journal. Authors should avoid making statements on clinical recommendations, economic benefits, and costs unless the manuscript includes the appropriate economic data and analyses (https://www.icmje.org/).

Phrasing statistical terms in manuscripts is known to be difficult. For instance, when authors report that 2 groups have different mean values, but the difference is not statistically significant, the observed difference is often described as *"there was no difference."* Another common phrase is that *"there was no statistical difference."* Whereas the first description is erroneous because it is a misrepresentation of what has been observed, the second is ridiculous because a statistically insignificant difference in mean values is as much "statistical" as a statistically significant one. Observed data should be described correctly (e.g., "the two groups had similar mean values" or *"differed in mean value, but the difference was not statistically significant"*). When preparing the manuscript, authors should consider both the practical effects (the clinical significance of the effects) and their estimation uncertainty (related to the statistical significance of the effects). Merely presenting a finding as "significant" is ambiguous.

While it is true that different analyses can give different results, this emphasizes the importance of planning and reporting your analyses and discuss all the options tested in the submitted research paper, not just the ones that were convincing. While research allows individual academics to pursue their interests, to learn something new, to hone their academic skills and to challenge themselves in new ways, research should be considered an endeavor built on systematic, honest investigation, seeking to expand the understanding of the world. Over the last decade, the concept of P hacking (and *"torturing of the data until it confesses"*) has made biomedical research journals aware of how authors use too many "valid statistical procedures" and then unfortunately end up selecting the one that leads to the most flattering conclusion. At *Acta Orthopaedica* we look forward to receiving many good papers, advocating that any departure from good basic premises (as outlined above) is not a valid practice of science and has the potential to do more harm than good by replacing truth and trustworthy investigation with shoddiness and falsehood.

Scientific progress is made in small incremental steps, over many years. It is a continuous process that takes time and

effort. Authors who get overly creative in interpreting spurious *post hoc* findings do nothing to add value to the total body of evidence. Rather, authors should practice careful study design, properly collect and handle data, avoid bias, and provide an honest representation of what was found without adding *"spin"* to their scientific manuscripts which includes the Discussion section.

Reporting the Discussion

We find it useful to begin the Discussion section by briefly summarizing the main findings based on the aim and hypotheses. Possible mechanisms or explanations for them are discussed and elaborated on and should be based on good scientific judgement. Authors should aim to emphasize the new and important aspects of their study and put their findings in the context of the totality of the relevant evidence. The editors of *Acta Orthopaedica* encourage authors to include a clear summary of previous research findings and to explain how their findings affect this summary. Authors should set the new results in the context of updated evidence balancing both previous and the new findings, thereby showing what contribution the new study has made to the cumulated evidence. Authors are cautioned to avoid being zealous or overly enthusiastic when interpreting findings. Rather, they should practice extra caution when interpreting potentially spurious findings, clarifying whether the new data supports (or refutes) their prior (prespecified) hypotheses.

Authors should state the potential limitations of their study. If authors imply what their research could lead to in clinical practice or policy, they must also be cautious and cognizant of the conclusions some readers might arrive at. The authors should link the conclusions with the (original) aim of the study while being careful not to provide unqualified statements that are not adequately supported by data; for instance, even serendipitous findings from exploratory analyses should be reported as such (i.e., not being based on an *a priori* hypothesis).