Statistical errors in medical research – a review of common pitfalls

Alexander M. Strasak, Qamruz Zaman, Karl P. Pfeiffer, Georg Göbel, Hanno Ulmer
Department of Medical Statistics, Informatics and Health Economics, Medical University Innsbruck

Summary

Background: standards in the use of statistics in medical research are generally low. A growing body of literature points to persistent statistical errors, flaws and deficiencies in most medical journals.

Methods: in this paper we present a comprehensive review of common statistical pitfalls which can occur at different stages in the scientific research process, ranging from planning a study, through conducting statistical data analysis and documenting statistical methods applied, to the presentation of study data and interpretation of study results.

Results: 47 potential statistical errors and shortcomings, differentiated for the distinct phases of medical research are presented and discussed.

Conclusions: statisticians should be involved early in study design, as mistakes at this point can have major repercussions, negatively affecting all subsequent stages of medical research. Consideration of issues discussed in this paper, when planning, conducting and preparing medical research manuscripts, should help further enhance statistical quality in medical journals.

Key words: statistics in medical research; common pitfalls and errors; study design; statistical analysis; documentation; presentation; interpretation

Introduction

Today, statistics is widely accepted as a powerful tool in the scientific research process with a great increase in the use of statistical methods having been documented for a wide range of medical journals over the past four decades [1–3]. Nevertheless there is wide consensus that standards are generally low, as a large proportion of published medical research contains statistical errors and shortcomings [4–12]. The problem is a serious one, as the inappropriate use of statistical analysis may lead to incorrect conclusions, artificial research results and a waste of valuable resources.

The misuse of statistics in medical research has therefore been widely discussed, and it has been pointed out that it is both unethical and can have serious clinical consequences [13, 14]. As a result, valuable efforts have been made by many medical journal editors to enhance the quality of statistics by adopting statistical guidelines for authors or by sharpening the statistical peer reviewing of incoming manuscripts [15–19]. Despite these efforts, there is little evidence that standards have improved over time, with recent studies pointing towards a persistence of major problems [11, 12, 20, 21].

In this paper we present a comprehensive review of common statistical errors, flaws and pitfalls concerning different stages of the medical scientific research process, from the planning of a medical study to the preparation of the final manuscript. The issues discussed are intended to help medical researchers to focus on what is important statistically and present it properly in their research papers. We have not set out to provide a set of fixed rules which should be followed, but rather give some advice and information about general statistical problems concerning aspects of study design, data analysis, documentation of statistical methods applied, presentation of study data and interpretation of study results, which can then be easily avoided, in the hope that statistical quality of medical research publications can be improved.

Each section of the paper first discusses the major statistical issues, which should be considered in the various stages of conducting medical research, followed by a summary of important statistical errors, flaws and pitfalls to be avoided. Section 2 draws attention to major statistical aspects regarding the design and planning of a study and points out some frequent errors and shortcomings, often occurring in this initial stage of a research project. Sections 3 and 4 deal with issues concern-
ing statistical data analysis and documentation of statistical methods employed, respectively, whereas sections 5 and 6 are devoted to statistical aspects regarding the proper and meaningful presentation of study data and interpretation of study results.

Study design

The most important phase of any research is the planning and design phase, as a proper and complete study design constitutes the basis for healthy research [1–3]. Errors, flaws and shortcomings, occurring in the planning stage can have a vast negative impact on the validity and reliability of research results, as they affect all subsequent stages of an investigation.

First of all, when planning a research project, it is important that the study aims, primary, as well as secondary outcome measures and study endpoints are formulated properly in the initial study protocol and also included adequately in the final research manuscript. The statistical as well as scientific hypotheses under investigation should be pre-specified and explicitly mentioned, especially if not self-evident, or if more than one hypothesis is being tested [22, 23]. If no pre-specified hypotheses are investigated, the exploratory character of the study should be outlined adequately.

It is crucial for high quality statistical work, to consider a-priori effect and sample size estimation and to appropriately conduct a statistical power calculation in the planning stage, to make sure that a study is provided with sufficient statistical power to detect treatment effects under observation and to avoid possible type II errors [1–3, 20, 24–26]. Further, the sample size used should always be stated in the manuscript and potential withdrawals during any phase have to be recorded in the study protocol and research manuscript.

Whenever possible, random samples, randomisation and blinding should be used to avoid potential bias, although this “gold standard” is not applicable for all types of studies and research projects [27]. Nevertheless, one has to be very aware that all inferential statistical techniques are valid only for random samples, and do not necessarily hold for data collected “any old how”. A full explanation of the method of randomisation and sampling should be mandatory when composing a research manuscript [5, 22].

If control groups are used, the initial equality and comparability of study groups has to be proven at baseline, to ensure that one does not use partially or inherently heterogeneous material, which is not comparable [22, 24]. In this case, possible differences after intervention can not be assigned logically to the treatment under investigation, unless multivariable techniques are applied adequately to adjust a study for these confounding factors. Nevertheless, applying statistical significance tests for comparing baseline balance of study groups (a common practice), is neither necessarily adequate nor advisable [28]. Merely by showing that there are no statistically significant differences between study groups at baseline, it can not be concluded that groups are equivalent, particularly in studies with small sample size that lack statistical power. One can only estimate the probability of obtaining the data given the truth of the null hypothesis, not the other way round [25]. Again, what is really needed is covariate adjustment for important confounders. Table 1 summarises important statistical errors and pitfalls related to the planning and design of a study, to help medical researchers to avoid them in their research.

<table>
<thead>
<tr>
<th>Study aims and primary outcome measures not clearly stated or unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to report number of participants or observations (sample size)</td>
</tr>
<tr>
<td>Failure to report withdrawals from the study</td>
</tr>
<tr>
<td>No a priori sample size calculation/effect-size estimation (power calculation)</td>
</tr>
<tr>
<td>No clear a priori statement or description of the Null-Hypothesis under investigation</td>
</tr>
<tr>
<td>Failure to use and report randomisation</td>
</tr>
<tr>
<td>Method of randomisation not clearly stated</td>
</tr>
<tr>
<td>Failure to use and report blinding if possible</td>
</tr>
<tr>
<td>Failure to report initial equality of baseline characteristics and comparability of study groups</td>
</tr>
<tr>
<td><strong>Use of an inappropriate control group</strong></td>
</tr>
<tr>
<td>Inappropriate testing for equality of baseline characteristics</td>
</tr>
</tbody>
</table>
Data analysis

When performing statistical data analysis and applying statistical significance tests or estimation techniques, it should be clear in one's mind that each method is based on several underlying assumptions, which have, at least approximately, to be fulfilled in order to obtain correct and meaningful results. Unfortunately, even simple and basic procedures such as the popular t-tests or chi-square tests are frequently misused in medical research, because their test assumptions are not evaluated sufficiently before application [5, 9, 29].

Moreover, when applying t-tests or chi-square tests, one has to be aware of choosing the correct version of the test, as they exist in various forms. If expected counts in a cell are less than 5, chi-square techniques should preferably be avoided, as their approximation under these circumstances is no longer reliable. If the sample is small, a Yates-continuity-correction should be employed, but preferably, exact tests should be used, in order to obtain reliable test results. Furthermore, the huge variety of statistical techniques now available means the choice of the most suitable and powerful one is not always trivial, since many details have to be considered.

If a study contains multiple endpoints necessitating multiple statistical tests, it is important to check the rate of false positive results and the potential inflation of type I errors by applying adequate multiple-comparison corrections [25, 29, 30]. Especially important is the recognition that comparing more than two groups, demands usage of methods of parametric or nonparametric analysis of variance, as repeated application of two group tests also boosts the risk of false positive test results. However, as multiple testing often occurs as a result of poor study design, it can easily be avoided by consulting a statistician early in the planning stage of a study.

Post-hoc subgroup analysis, not pre-specified in the initial study protocol should also be avoided, as it leads to the impression of “shopping” for statistically significant results [31]. Moreover, application of multivariate techniques should be mandatory if confounding factors are assumed to be present or if study groups are not adequately matched for relevant baseline characteristics. Table 2 summarises major statistical errors and deficiencies related to data analysis, to help medical researchers to avoid them in their own research.

Table 2
Statistical errors and deficiencies related to data analysis.

<table>
<thead>
<tr>
<th>Use of wrong statistical tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incompatibility of statistical test with type of data examined</td>
</tr>
<tr>
<td>Unpaired tests for paired data or vice versa</td>
</tr>
<tr>
<td>Inappropriate use of parametric methods</td>
</tr>
<tr>
<td>Use of an inappropriate test for the hypothesis under investigation</td>
</tr>
<tr>
<td>Inflation of Type I error</td>
</tr>
<tr>
<td>Failure to include a multiple-comparison correction</td>
</tr>
<tr>
<td>Inappropriate post-hoc Subgroup analysis</td>
</tr>
</tbody>
</table>

Typical errors with Student’s t-test

| Failure to prove test assumptions |
| Unequal sample sizes for paired t-test |
| Improper multiple pair-wise comparisons of more than two groups |
| Use of an unpaired t-test for paired data or vice versa |

Typical errors with χ²-tests

| No Yates-continuity correction reported if small numbers |
| Use of chi-square when expected numbers in a cell are <5 |
| No explicit statement of the tested Null-Hypotheses |
| Failure to use multivariate techniques to adjust for confounding factors |

Documentation

It is essential to sound scientific practise, that all statistical methods applied are described clearly, correctly and with enough detail, to enable a knowledgeable reader with access to the study data, to recalculate all results. Therefore, a subsec-

tion devoted exclusively to concerns of statistical analysis, where all techniques and methods used are mentioned, is mandatory in every medical research paper.

Commonly used methods do not need to be explained in detail, while any new application and reasons for using the method should be summarised or referenced [1–3]. When using more than one statistical test, it is important to specify which test was performed on a given set of data [9, 20, 32]. The notion of a simple “where appropriate” statement [32], although indicating the good intentions of the authors in applying the correct tests, is not sufficient in this context.

For statistical tests, which exist as paired or unpaired versions (eg, t-test, Wilcoxon-test), it is es-

Table 3
Errors related to the documentation of statistical methods applied.

| Failure to specify/define all tests used clear and correctly |
| Failure to state number of tails |
| Failure to state if test was paired or unpaired |
| Wrong names for statistical tests |
| Referring to unusual or obscure methods without explanation or reference |
| Failure to specify which test was applied on a given set of data if more than one test was done |
| “Where appropriate” statement |
According to Evans [33] good research deserves to be well presented and sound presentation is as much a part of the research as the collection and analysis of the data. Therefore, when statistically describing or presenting study data, one should be aware of using adequate statistical measures of central tendency and dispersion. If using arithmetic means and standard deviations, it should be evident, that the data is at least approximately normally distributed and not skewed. Otherwise these measures can not be used meaningfully for describing the data. In every case, standard deviations should preferably be reported in parentheses [ie, mean (SD)] than using mean ± SD expressions, as the latter specification can be confused with a 95% confidence interval by the reader. For skewed data, as often the case in biological and medical research, giving medians, quartiles or ranges is more suitable, although one has to be aware that the range is sensitive to outliers and hence sometimes may be unsuitable as a summary statistic. If consecutively applying nonparametric tests for statistical data analysis, one should avoid giving means and standard deviations, as these parameters are, by definition, not tested by a nonparametric test and hence do not make sense for describing the data under investigation. Here medians, ranges or interquartile-ranges should have preference. It is not sufficient, however, to just present mean values without giving any measures of variability of the data [4, 22, 23, 29].

The standard error of the mean (SEM), although commonly and erroneously used for statistical description (perhaps because it makes data seem less variable), is not a descriptive statistic, but rather an inferential method used for statistical estimation [34–36]. Thus, it is not appropriate to present a mean together with a standard error of the mean as a measure of dispersion. In the same vein, it is essential that any “±” notion used in the text, in charts, tables or figures is specified when it first appears, and an adequate description is given of the meaning of error bars in graphical illustrations [37].

For primary endpoints and main outcome measures of a study, confidence intervals should be estimated whenever possible, as a probability value alone does not give meaningful information about the magnitude or size of an effect [1–3, 24]. Thus the proper application of techniques of statistical estimation can largely extend the information content of a study for its readers. If using statistical estimation for comparing groups, confidence intervals should rather be given for the differences between groups, than for each group itself.

P-values should be reported as exactly obtained, rather than with arbitrarily allocated thresholds for example, “p = ns”, “p <0.05”, or “p >0.05”. However, numerical information should not be given to an unrealistic level of precision, which cannot be justified by the study’s sample size.

Table 4 summarises statistical errors and flaws related to the presentation of study data.

<table>
<thead>
<tr>
<th>Inadequate graphical or numerical description of basic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean but no indication of variability of the data</td>
</tr>
<tr>
<td>Giving SE instead of SD to describe data</td>
</tr>
<tr>
<td>Use of mean (SD) to describe non-normal data</td>
</tr>
<tr>
<td>Failure to define ± notion for describing variability or use of unlabeled error bars</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inappropriate and poor reporting of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results given only as p-values, no confidence intervals given</td>
</tr>
<tr>
<td>Confidence intervals given for each group rather than for contrasts</td>
</tr>
<tr>
<td>&quot;p = NS&quot;, &quot;p &lt;0.05&quot; or other arbitrary thresholds instead of reporting exact p-values</td>
</tr>
<tr>
<td>Numerical information given to an unrealistic level of precision</td>
</tr>
</tbody>
</table>

Interpretation

In the final stage of the scientific research process, when interpreting results of data analysis, it is essential that the most suitable and powerful statistical tests were applied. This avoids drawing conclusions from a study, which are insufficiently supported by the data. If claiming significance of effects, one has to ensure, that a statistical significance test has been employed. If results do not exhibit statistical significance, it is crucial to be careful in drawing conclusions, as lack of statistical significance does not invariably mean there was no effect or no difference at all [20, 23, 24]. For example, perhaps merely the sample size was too small to safeguard statistical significance, although the
Table 5
Statistical errors and deficiencies related to the interpretation of study findings.

<table>
<thead>
<tr>
<th>Wrong interpretation of results</th>
<th>Poor interpretation of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>“non significant” interpreted as “no effect”, or “no difference”</td>
<td>Disregard for Type II error when reporting non-significant results</td>
</tr>
<tr>
<td>Drawing conclusions not supported by the study data</td>
<td>Missing discussion of the problem of multiple significance testing if done</td>
</tr>
<tr>
<td>Significance claimed without data analysis or statistical test mentioned</td>
<td>Failure to discuss sources of potential bias and confounding factors</td>
</tr>
</tbody>
</table>

Discussion

The aim of the paper at hand is to comprehensively review frequently observed statistical errors, flaws and pitfalls in medical science in order to help medical researchers produce statistically sound output in their future investigations. Applying these principles should lead to an improvement in the statistical quality of research papers published in medical journals. As the issues discussed in this paper largely concentrate on simple but nevertheless important statistical matters, they can easily be handled and particularly aid the non-statistician in conducting statistically robust research, without needing to read whole textbooks on statistical methodology.

However, for a long-term rise in standards in the use of statistics in medical research to occur, it is necessary to refer to and access a broader range of sources than those given in this paper. Additional aspects concerning broader methodological and statistical issues in randomised controlled trials, meta-analyses and diagnostic studies, for example, are illustratively summarised in the Consort statement [38], the Quorom statement [39] and the Stard statement [40], which all contain a wealth of useful information.

Although the availability of multifaceted statistical software packages makes it easy for statistically unskilled clinicians to conduct their own data analysis, this can lead to major problems arising from insufficient knowledge of the underlying mathematical concepts or statistical ideas. Medical researchers have to be encouraged to learn more about statistics, as various studies point to a lack of statistical knowledge among medical residents [41, 42]. Furthermore, consulting a statistician and interaction with skilled experts often takes place when the planning and designing of a study is long done. This makes it difficult to straighten out statistical shortcomings, which occurred in earlier steps of an investigation [43, 44].

Results possibly still contain clinically important findings or give a worthwhile impulse for fellow researchers.

When using small sample sizes and obtaining non-significant test results, a discussion of possible type II errors should be obligatory [9, 26], as well as sufficient mention of the problem of multiple significance testing and the increasing risk of false positive test results (if these occurred) [23]. If a study is not adjusted for potential confounding factors or bias, this also should be adequately discussed when interpreting the results. Table 5 summarises important statistical errors and pitfalls related to the interpretation of study findings.

Medical journal editors should seriously consider broader enhancement of quality by intensifying the statistical reviewing of incoming manuscripts, as there is also strong evidence, that the process of statistical peer reviewing of medical journals is insufficient and ineffective [45]. In this context it is desirable for all papers to be reviewed by a statistician prior to publication and journals should publish their policy on statistical reviewing. Statistical reviewers should at least be offered the option to see the revised manuscripts before final publication [2].

The careful and accurate use of statistics in medical research is of major importance and therefore must be enforced emphatically. The use of statistics in medical diagnoses and biomedical research may affect whether individuals live or die, whether their health is protected or jeopardised, and whether medical science advances or gets sidetracked. [...] Because society depends on sound statistical practice, all practitioners of statistics, whatever their training and occupation, have social obligations to perform their work in a professional, competent, and ethical manner.” [Ethical Guidelines for Statistical Practise, American Statistical Association, 1999]. We hope that this paper will contribute to a step in the right direction.

Correspondence:
Dr. Alexander M. Strasak
Department of Medical Statistics, Informatics and Health Economics
Medical University Innsbruck
Schoepfstrasse 41
A-6020 Innsbruck
E-Mail: alexander.strasak@i-med.ac.at
References

The many reasons why you should choose SMW to publish your research

**What Swiss Medical Weekly has to offer:**

- SMW’s impact factor has been steadily rising. The 2005 impact factor is 1.226.
- Open access to the publication via the Internet, therefore wide audience and impact.
- Rapid listing in Medline.
- LinkOut-button from PubMed with link to the full text website http://www.smw.ch (direct link from each SMW record in PubMed).
- No-nonsense submission – you submit a single copy of your manuscript by e-mail attachment.
- Peer review based on a broad spectrum of international academic referees.
- Assistance of our professional statistician for every article with statistical analyses.
- Fast peer review, by e-mail exchange with the referees.
- Prompt decisions based on weekly conferences of the Editorial Board.
- Prompt notification on the status of your manuscript by e-mail.
- Professional English copy editing.
- No page charges and attractive colour offprints at no extra cost.

**International Advisory Committee**

Prof. K. E. Juhani Airaksinen, Turku, Finland
Prof. Anthony Bayes de Luna, Barcelona, Spain
Prof. Hubert E. Blum, Freiburg, Germany
Prof. Walter E. Haefeli, Heidelberg, Germany
Prof. Nino Kuenzli, Los Angeles, USA
Prof. René Lutter, Amsterdam, The Netherlands
Prof. Claude Martin, Marseille, France
Prof. Josef Patsch, Innsbruck, Austria
Prof. Luigi Tavazzi, Pavia, Italy

We evaluate manuscripts of broad clinical interest from all specialities, including experimental medicine and clinical investigation.

We look forward to receiving your paper!

Guidelines for authors:
http://www.smw.ch/set_authors.html

**Editorial Board**

Prof. Jean-Michel Dayer, Geneva
Prof. Peter Gehr, Berne
Prof. André P. Perruchoud, Basel
Prof. Andreas Schaffner, Zurich (Editor in chief)
Prof. Werner Straub, Berne
Prof. Ludwig von Segesser, Lausanne

EMH Swiss Medical Publishers Ltd.
SMW Editorial Secretariat
Farnburgerstrasse 8
CH-4132 Muttenz

Manuscripts: submission@smw.ch
Letters to the editor: letters@smw.ch
Editorial Board: red@smw.ch
Internet: http://www.smw.ch

All manuscripts should be sent in electronic form, to: