What does it take from statistical point of view to get published in a high-impact journal?

Jonas Ranstam
Statistical consultant/reviewer/editor

- Acta Orthopaedica, 1993-
- British Journal of Surgery, 2014-
- Lancet, 2013-
- Lancet Diabetes and Endocrinology, 2013-
- Lancet Global Health, 2013-
- Lancet Neurology, 2013-
- Lancet Respiratory Medicine, 2013-
- Lancet Psychiatry, 2014-
- Osteoarthritis and Cartilage, 2008-
- Statistical Methods in Medical Research, 1999-2004
## Number of reviewed manuscripts

<table>
<thead>
<tr>
<th>Journal</th>
<th>Nr of reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acta Orthopaedica</td>
<td>1 442</td>
</tr>
<tr>
<td>British Journal of Surgery</td>
<td>5</td>
</tr>
<tr>
<td>The Lancet</td>
<td>30</td>
</tr>
<tr>
<td>The Lancet Diabetes Endocrinology</td>
<td>12</td>
</tr>
<tr>
<td>The Lancet Global Health</td>
<td>4</td>
</tr>
<tr>
<td>The Lancet Respiratory Medicine</td>
<td>5</td>
</tr>
<tr>
<td>The Lancet Neurology</td>
<td>5</td>
</tr>
<tr>
<td>Osteoarthritis and Cartilage</td>
<td>1 387</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>2 890</strong></td>
</tr>
</tbody>
</table>

1. Reviewed prior to 5 March, 2014 (excluding some 600-700 manuscripts in Acta Orthop's earlier, non-electronic, manuscript system).
The short answer: Nothing special
The long answer

What characterizes high impact journals?

- General vs. specialty journal
- News orientation
- Statistical reviewing
Plan

1. The impact factor
2. Manuscript handling
3. Statistical reviewing
4. Important guidelines
5. Statistical advisers
6. Discussion
1. The Impact factor
Impact Factor (IF)

The IF for a journal is calculated based on a 3-year period, and can be considered to be the average number of times published papers are cited up to 2 years after publication.

The impact factor 2014 for a journal would be calculated as follows:

\[ A = \text{the number of times articles published in 2012-2013 were cited in indexed journals during 2014} \]

\[ B = \text{the number of articles, reviews, proceedings or notes published in 2012-2013} \]

\[ \text{impact factor 2014} = \frac{A}{B} \]
# Impact factors 2013 (selected journals)

<table>
<thead>
<tr>
<th>Journal Name</th>
<th>Impact Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA: CANCER J CLIN</td>
<td>102</td>
</tr>
<tr>
<td>NEW ENGL J MED</td>
<td>53</td>
</tr>
<tr>
<td>LANCET</td>
<td>38</td>
</tr>
<tr>
<td>NATURE</td>
<td>36</td>
</tr>
<tr>
<td>NAT GENET</td>
<td>36</td>
</tr>
<tr>
<td>CELL</td>
<td>32</td>
</tr>
<tr>
<td>SCIENCE</td>
<td>31</td>
</tr>
<tr>
<td>JAMA</td>
<td>30</td>
</tr>
<tr>
<td>CANCER CELL</td>
<td>26</td>
</tr>
<tr>
<td>NAT IMMUNOL</td>
<td>26</td>
</tr>
<tr>
<td>LANCET NEUROL</td>
<td>23</td>
</tr>
<tr>
<td>NAT BIOTECHNOL</td>
<td>23</td>
</tr>
<tr>
<td>LANCET ONCOL</td>
<td>22</td>
</tr>
<tr>
<td>NAT MED</td>
<td>22</td>
</tr>
<tr>
<td>ANN INTERN MED</td>
<td>16</td>
</tr>
<tr>
<td>PLOS MED</td>
<td>16</td>
</tr>
<tr>
<td>CIRCULATION</td>
<td>14</td>
</tr>
<tr>
<td>BRIT MED J</td>
<td>14</td>
</tr>
<tr>
<td>BRIT J SURG</td>
<td>5</td>
</tr>
<tr>
<td>OSTEOARTH CARTILAGE</td>
<td>4</td>
</tr>
<tr>
<td>ACTA ORTHOP</td>
<td>2</td>
</tr>
<tr>
<td>RUSS J CARDIOL</td>
<td>0.005</td>
</tr>
</tbody>
</table>
The impact factor has been criticized

Journals covering a broad area of science with a rapidly growing but ephemeral literature that tends to cite many articles, get higher IF than specialist journals whose articles may reach peak citation many years after their publication.

Review articles tend to be much more heavily cited than primary research articles and hence journals carrying some review articles, and especially journals exclusively devoted to them, will have high IF.

The distribution of citations received by individual articles in a journal is so broad as to make the mean almost meaningless.
A valid measure of quality?

Objectives: The authors assessed the validity of impact factor as a measure of quality for general medical journals by testing its association with journal quality as rated by clinical practitioners and researchers.

Results: The correlation between impact factor and physicians' ratings of journal quality was strong ($r^2 = 0.82$, $P = 0.001$). The correlation was higher for the research group ($r^2 = 0.83$, $P = 0.001$) than for the practitioner group ($r^2 = 0.62$, $P = 0.01$).

Conclusions: Impact factor may be a reasonable indicator of quality for general medical journals.

2. Manuscript handling
Manuscript handling

The Lancet has an in-house team of 12 physicians and scientists who read around 8,000 submitted papers each year (not including the specialty journals).

About 70% of the manuscripts are rejected outright.

The remaining 30% are sent for external peer review.
Manuscript handling

After peer review, the manuscript is presented to colleagues at a once-weekly manuscript meeting.

All editors collectively make the decision whether to proceed to statistical review, to seek a revised paper from the authors, to seek further reviews, or to reject the paper.

All original research manuscripts published in The Lancet have at least 3 clinical and 1 statistical review.
Manuscript handling

The statistical review is usually a fast-track review

Review often to be returned within 48 hours (in some cases 5 days).
3. Statistical reviewing
What the editorial office expects from the statistical review

The Lancet

1. An overall decision on the manuscript (confidential)

2. Answers to the manuscript rating questions (confidential)

3. Comments to author (reviewer blinded in some journals)
   - A detailed appraisal for the author

4. Comments to editor (confidential)
   - A frank account of the strengths and weaknesses of the article
   - A declaration of conflicts of interest
Statistical reviewing

Common misunderstandings among authors

1. $P < 0.05$ indicates that a finding is “scientific”

2. $P > 0.05$ indicate similarity, “no difference”

3. Calculating a p-value is a practical problem (running a program)

4. Everyone can calculate a p-value with SPSS

5. A statistical reviewer checks that p-values are calculated correctly
Irreproducible results

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

“I SEE a train wreck looming,” warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned research on a phenomenon known as “priming”. Priming studies suggest that decisions can be influenced by apparently irrelevant actions or events that took place just before the cusp of choice. They have been a boom area in psychology over the past decade, and some of their insights have already made it out of the lab and into the toolkits of policy wonks keen on “nudging” the populace.
Scientists faces irreproducible results

RNA researcher defends experiments others have found impossible to repeat.

Ichiko Fuyuno

After the spectacular case of fraud involving stem-cell researcher Woo Suk Hwang, Asia has been hit by another, more low-key scandal.

The head of an investigating committee at the University of Tokyo announced on 27 January that at least one of the experiments performed by a Japanese RNA researcher, whose credibility stands...
Irreproducible results

No Cure
When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

- Fully replicated: 20.9%
- Partially replicated: 11.9%
- Not replicated: 64.2%
- Not applicable: 3.0%

Source: Nature Reviews Drug Discovery
Why Most Published Research Findings Are False
John P.A. Ioannidis

Summary
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the role of true or no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller, when study sizes are smaller, when there is greater number and lesser preselection of tested relationships, when there is greater flexibility in designs, definitions, outcomes, and analytical modes, when there is greater financial and other interest and prejudice, and when more teams are involved in a specific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias in the field. We discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Rebuttal and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–5] to the most modern molecular research [6,7]. There is increasing concern that in modern research, false findings may be the majority, even the vast majority of published research claims [8–11]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings
Several methodologies have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet discredited strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings. As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10, 11]. Consider a 2 x 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The post-study probability of a relationship being true is \( \frac{R}{R + 1} \). The probability of a study finding a true relationship reflects the power 1 – \( \beta \) (one minus the Type II error rate). The probability of claiming a relationship when some truth exists reflects the Type I error rate, \( \alpha \). Assuming that relationships are being probed in the field, the expected values of the 2 x 2 table are shown in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV.

The PPV is also the complementary probability of what Warholer et al. have called the false positive report probability [12]. According to the 2 x 2 table, one gets PPV = \( \frac{1 – \beta}{R/(R + 1)} \). A research finding is thus

Copyright: © 2005 John P.A. Ioannidis. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviation: PPV, positive predictive value

John P.A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research, and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: johni@cc.uoc.gr

Competing interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.002024
F or a brief moment in 2014, Matt Motyl was one of the brights of scientific glory. He had discovered that extremists quite literally see the world in black and white. The results were "plain to see," as he recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was very strong," he says, "and the data provided clear support." The *P* value, a common index of the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Amidst the controversy over reproducibility, Motyl and his advisor, Brian Nosek, decided to replicate the study. When the data, the *P* value came out as 0.39 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame.

It turned out that the problem was not in the dataset in Motyl's analysis. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an econometrician at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2013, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

At the same time, statisticians are looking for better ways of thinking about data, to help scientists avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Andrew Goodman, a physicist and statistician at Stanford. "There is an error handed down from God and no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

**OUT OF CONTEXT**

*P* values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes: annoying and impossible to swat away, the emperor's new clothes (adorned with obvious problems that everyone ignores) and the tool of a "likely intellectual rascal" who ravishes science but leaves it with no progeny. One recent paper suggests deconstructing the methodology "statistically by hypothesis inference testing," and reassessing the acronym it would yield.

The irony is that when U.K. statistician Ronald Fisher introduced the *P* value in the 1920's, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the...
Good statistics is a fundamental requirement

Among themselves, statisticians sometimes view their contribution to research in terms of a paraphrase of chemical giant BASF’s classic advertising tag line: “We don’t make the products. We make them better.” In doing so, they sell themselves short.

Good statistics can no longer be seen as something that makes science better — it is a fundamental requirement, and one that can only grow in importance as funding cuts bite and competition for resources intensifies.

*Nature 2014;506:131*
What makes a manuscript good from a statistical viewpoint?

“Statistics is the study of uncertainty”

*Dennis Lindley*
What makes a manuscript good from a statistical viewpoint?

“Statistical science is not just about the study of uncertainty but rather deals with inferences about scientific theories from uncertain data.”

John Nelder
A depressing observation

Many researchers do not grasp the duality of sample and population, the difference between observing and inferring.

They don't know what p-values measure, and they believe that confidence intervals describe data dispersion.
What statistical reviewing is and isn't

Statistical reviewing is not about checking compliance to rules of thumb.

Statistical reviewing is about evaluating whether the empirical support for conclusions presented in a manuscript is sufficient, consistent, and clearly described.
The responsibilities of a statistical reviewer

“To make sure that the authors spell out for the reader the limitations imposed upon the conclusions by the design of the study, the collection of data, and the analyses performed.”

Statistical reviewing

General issues

- Study aim
- Sample
- Data collection
- Statistical methods
- Fulfilment of underlying methodological assumptions
- Results (hypothesis test result, estimation uncertainty)
- Caveats (uncertainties, limitations)
- Conclusions
Statistical reviewing

Special issues

- Systematic reviews
- Randomized trials
- Observational studies
- Laboratory experiments
- Surveys
Statistical reviewing

Systematic reviews

- PRISMA statement
- Heterogeneity (fixed/random effects)
- Individual participant data (preserved clustering)
- Publication bias (funnel plots)
- Etc.
Statistical reviewing

Randomized trials

- CONSORT statement
- clinicaltrials.gov
- Helsinki Declaration
- Study protocol (check primary endpoint)
- ICH-GCP (esp. for drug trials)
- ITT/PP (superiority/non-inferiority)
- Endpoints
- Randomization
- Stratification (covariates)
- Multiplicity
- Etc.
Statistical reviewing

Observational studies

- Risk factor estimation - prediction
- Confounding by association
- Confounding by indication
- Misclassification/measurement errors
- Parameter interpretation (OR/RR)
- Etc.
Statistical reviewing

Laboratory experiments

- ARRIVE guidelines
- Pseudoreplication, pooled datasets, etc.
- Multiplicity issues (Bonferroni within endpoints)
- Etc.
Statistical reviewing

Surveys

- Finite populations
- Cluster sampling
- Scaled/weighted analyses
- Etc.
Statistical reviewing

Overall consideration

- Transparency and consistency from aim to conclusion.
4. Important guidelines
Important guidelines

Preparing for Submission

PAGE CONTENTS

1. General Principles
2. Reporting Guidelines
3. Manuscript Sections
   a. Title Page
   b. Abstract
   c. Introduction
   d. Methods
   e. Results
   f. Discussion
   g. References
   h. Tables
   i. Illustrations (Figures)
   j. Units of Measurement
   k. Abbreviations and Symbols
Important guidelines

The CONSORT Statement

The CONSORT Statement is intended to improve the reporting of a randomized controlled trial (RCT), enabling readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results. It emphasizes that this can only be achieved through complete transparency from authors.

Investigators and editors developed and revised the CONSORT (CONsolidated Standards of Reporting Trials) Statement to help authors improve reporting of two-parallel design RCTs by using a checklist and flow diagram. The most up-to-date revision of the CONSORT Statement is CONSORT 2010, which can be freely viewed and downloaded from this website. All previous versions of the CONSORT Statement are outdated.

Extensions of the CONSORT Statement have been developed for other types of study designs, interventions and data.

The Checklist

The checklist items pertain to the content of the Title, Abstract, Introduction, Methods, Results, Discussion, and Citing CONSORT.
## Important guidelines

### Annals of internal medicine

#### Information for Authors

#### Statistical Guidelines

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Percentages</strong></td>
<td>Report percentages to one decimal place (i.e., xx.%) when sample size is &gt;=200. To avoid the appearance of a level of precision that is not present with small samples, do not use decimal places (i.e., xx%, not xx.xxx%) when sample size is &lt; 200.</td>
</tr>
<tr>
<td><strong>Standard deviations</strong></td>
<td>Use “mean (SD)” rather than “mean ± SD” notation. The ± symbol is ambiguous and can represent standard deviation or standard error.</td>
</tr>
<tr>
<td><strong>Standard errors</strong></td>
<td>Report confidence intervals, rather than standard errors, when possible.</td>
</tr>
<tr>
<td><strong>P values</strong></td>
<td>For P values between 0.001 and 0.20, please report the value to the nearest thousandth. For P values greater than 0.20, please report the value to the nearest hundredth. For P values less than 0.001, report as “P&lt;0.001.”</td>
</tr>
<tr>
<td><strong>Trend</strong></td>
<td>Use the word trend when describing a test for trend or dose-response. Avoid the term trend when referring to P values near but not below 0.05. In such instances, simply report a difference and the confidence interval of the difference (if appropriate) with or without the P value.</td>
</tr>
<tr>
<td><strong>Statistical software</strong></td>
<td>Specify in the statistical analysis section the statistical software—version, manufacturer, manufacturer’s location, and the specific functions, procedures, or programs—used for analyses.</td>
</tr>
<tr>
<td><strong>Cox models</strong></td>
<td>When reporting the findings from Cox proportional hazards models:</td>
</tr>
<tr>
<td></td>
<td>Do not describe hazard ratios as relative risks.</td>
</tr>
<tr>
<td></td>
<td>Do report how the assumption of proportional hazards was tested, and what the test showed.</td>
</tr>
<tr>
<td><strong>Descriptive tables</strong></td>
<td>In tables that simply describe characteristics of 2 or more groups (e.g. Table 1 of a clinical trial):</td>
</tr>
<tr>
<td></td>
<td>Report averages with standard deviations, not standard errors, when data are normally distributed.</td>
</tr>
<tr>
<td></td>
<td>Report median (minimum, maximum) or median (25th, 75th percentile [interquartile range, or IQR]) when data are not normally distributed.</td>
</tr>
<tr>
<td></td>
<td>Avoid reporting P values as there can be imbalance when P’s are not significant (because of small sample size) and balance when P values are significant (because of large sample size).</td>
</tr>
</tbody>
</table>
Important guidelines

New England Journal of Medicine

Statistical Methods

- Exact methods should be used as extensively as possible in the analysis of categorical data. For analysis of measurements, nonparametric methods should be used to compare groups when the distribution of the dependent variable is not normal.
- Results should be presented with only as much precision as is of scientific value. For example, measures of association, such as odds ratios, should ordinarily be reported to two significant digits.
- Measures of uncertainty, such as confidence intervals, should be used consistently, including in figures that present aggregated results.
- Except when one-sided tests are required by study design, such as in noninferiority trials, all reported P values should be two-sided. In general, P values larger than 0.01 should be reported to two decimal places, those between 0.01 and 0.001 to three decimal places; P values smaller than 0.001 should be reported as P<0.001. Notable exceptions to this policy include P values arising in the application of stopping rules to the analysis of clinical trials and genetic-screening studies.
- For tables comparing treatment or exposure groups in a randomized trial (usually the first table in the trial report), significant differences between or among groups should be indicated by * for P < 0.05, ** for P < 0.01, and *** for P < 0.001 with an explanation in the footnote if required. The body of the table should not include a column of P values.
- In manuscripts that report on randomized clinical trials, authors may provide a flow diagram in CONSORT format and all of the information required by the CONSORT checklist. When restrictions on length prevent the inclusion of some of this information in the manuscript, it may be provided in a separate document submitted with the manuscript. The CONSORT statement, checklist, and flow diagram are available on the CONSORT website.
Important guidelines

Editorials and similar


5. About statistical advisers
About statistical advisers

Prevalence

Most medical journals (70%-80% of Elsevier's) have one or more statistical advisers

Br J Surg has 2

The Lancet has 28
About statistical advisers

Assignment

- Assigned by the editor-in-chief
- Elected by the editorial board
- Recruited by the publisher
About statistical advisers

Typical (guesstimated) work load and payment

3h - 5h per manuscript
$100 - £100 per manuscript
50 - 100 manuscripts per year and journal
6. Discussion
Thank you for your attention!