ELSEVIER

CrossMark

Journal of Clinical Epidemiology 87 (2017) 70–77

Journal of Clinical Epidemiology

The distribution of *P*-values in medical research articles suggested selective reporting associated with statistical significance

Thomas V. Perneger*, Christophe Combescure

Division of Clinical Epidemiology, Faculty of Medicine, University of Geneva, and Geneva University Hospitals, 6 rue Gabrielle-Perret-Gentil, CH-1211 Geneva, Switzerland

Accepted 4 April 2017; Published online 9 April 2017

Abstract

Objectives: Published *P*-values provide a window into the global enterprise of medical research. The aim of this study was to use the distribution of published *P*-values to estimate the relative frequencies of null and alternative hypotheses and to seek irregularities suggestive of publication bias.

Study Design and Setting: This cross-sectional study included *P*-values published in 120 medical research articles in 2016 (30 each from the *BMJ*, *JAMA*, *Lancet*, and *New England Journal of Medicine*). The observed distribution of *P*-values was compared with expected distributions under the null hypothesis (i.e., uniform between 0 and 1) and the alternative hypothesis (strictly decreasing from 0 to 1). *P*-values were categorized according to conventional levels of statistical significance and in one-percent intervals.

Results: Among 4,158 recorded *P*-values, 26.1% were highly significant ($P < 0.001$), 9.1% were moderately significant ($P \geq 0.001$ to < 0.01), 11.7% were weakly significant ($P \geq 0.01$ to < 0.05), and 53.2% were nonsignificant ($P \geq 0.05$). We noted three irregularities: (1) high proportion of *P*-values <0.001, especially in observational studies, (2) excess of *P*-values equal to 1, and (3) about twice as many *P*-values less than 0.05 compared with those more than 0.05. The latter finding was seen in both randomized trials and observational studies, and in most types of analyses, excepting heterogeneity tests and interaction tests. Under plausible assumptions, we estimate that about half of the tested hypotheses were null and the other half were alternative.

Conclusion: This analysis suggests that statistical tests published in medical journals are not a random sample of null and alternative hypotheses but that selective reporting is prevalent. In particular, significant results are about twice as likely to be reported as nonsignificant results.

Keywords: Statistical tests; *P*-values; Publication bias; Practice of research

1. Introduction

Most medical research studies, regardless of design or purpose, report results accompanied by *P*-values or by confidence intervals [1]. The aggregate population of published *P*-values (or confidence intervals) can be seen as a collective artifact of the medical research enterprise that may reveal useful clues about the conduct of science and the dissemination of scientific results.

The main issue that hampers the empirical study of *P*-values is selection bias [2]. This bias can occur both through the researcher's ingenuity in finding a "statistically significant" result (a practice sometimes called "*P*-hacking" [3]) and from preferential publication of significant results [4], attributable to both researchers and journal editors. Recent studies found an unusually high occurrence of *P*-values just below the threshold of statistical significance [5–7]. For example, in abstracts that reported results as odds ratios or relative risks, Gøtzsche found 46 *P*-values between 0.0400 and 0.0499, but only five between 0.0500 and 0.0599 [5]; which would be highly unlikely without selection bias. When Jager and Leek estimated the "science-wise false discovery rate" (i.e., the proportion of published significant findings that correspond to type-1 errors) by applying statistical models developed for genomic studies [8], their approach was criticized chiefly because publication bias renders the statistical model untrustworthy [9–12].

Previous studies have not fully reflected what happens in medical research because they examined only abstracts

Conflict of interest: None.

* Corresponding author.
E-mail address: thomas.perneger@hcuge.ch (T.V. Perneger).

http://dx.doi.org/10.1016/j.jclinepi.2017.04.003
0895-4356/

What is new?

Key findings

- The distribution of >4,000 *P*-values published in medical research articles suggested a pervasive selection bias associated with statistical significance.
- This bias was observed for most study designs and most types of analyses, including randomized trials and primary analyses but excepting interaction tests and heterogeneity tests.

What this adds to what was known?

- Previous studies have shown that *P*-values published in abstracts are highly selected to highlight results that are statistically significant.
- This study suggests that selective reporting of *P*-values affects medical research articles globally.

What is the implication and what should change now?

- The focus on statistical significance distorts the published record of medical research and the evidence available for medical decision making.
- Other methods of statistical inference than *P*-values and other methods for disseminating research results deserve consideration.

[5,7,8] or only a subset of *P*-values [6]. In this study, we describe the distribution of *P*-values in full medical research articles, to verify if this distribution matches the shape that would be expected from a mixture of null and alternative hypotheses, and to identify irregularities that may reveal selection bias. We compare distributions of *P*-values according to study design and type of analysis. In this study, we observe scientific practice but do not attempt to judge the appropriateness of the statistical tests that were performed, nor the adequacy of their interpretation. Neither do we address the fundamental merits and limitations of *P*-values as measures of evidence; this issue is addressed elsewhere (e.g., [13,14]).

2. Methods

We included in this cross-sectional study medical research articles published in four prominent journals (*BMJ*, *JAMA*, *Lancet*, and *New England Journal of Medicine*) starting on April 1, 2016. We identified articles that analyzed original numerical data and included at least one *P*-value, in order of publication, until 30 eligible articles were retrieved from each journal (120 articles in total). We also noted the usage of estimation methods (typically confidence intervals, in a few cases credible intervals), either alone or in conjunction with *P*-values, in all screened articles.

All reported *P*-values were retrieved from the selected articles, as published, except when the result was given as significant or not at the 0.05 level or described verbally as such. We did not retrieve *P*-values from appendices or attachments. For each article, we first abstracted *P*-values from the tables (including footnotes), then from the figures, and finally from the text, making sure to skip *P*-values that replicated those from tables or figures. We identified for each *P*-value the following information: (1) appearance in the abstract, (2) whether it was a primary analysis (according to the Methods section of each article), (3) whether it came from a parsimonious model or was otherwise described as a significant result selected among a larger number of results, (4) whether it was a baseline comparison from a randomized trial, (5) whether it was an interaction test that was not the primary analysis, (6) whether it was a heterogeneity test from a meta-analysis, (7) whether any correction for multiple testing had been applied. In the latter case, we did not back-compute uncorrected *P*-values as insufficient detail was provided in all instances.

The sample size for this study was chosen so as to obtain a sufficiently detailed description of the distribution of *P*-values. We aimed to obtain at least 20 *P*-values in each one-percent interval; because we expected the distribution to be skewed to the right, we decided to retrieve about 4,000 *P*-values in total. As we also wanted to include the same number of articles from each journal, the final number of *P*-values was 4,158.

For the analysis, *P*-values smaller than 0.01 reported as inequalities were imputed to the midpoint of the corresponding interval, for example, for $P < 0.001$, we imputed 0.0005. Based on the initial exploratory analysis which identified irregular frequencies at both extremities of the distribution and at $P = 0.05$, we classified the *P*-values into six categories: (1) <0.001, (2) ≥ 0.001 to <0.01, (3) ≥ 0.01 to <0.05, (4) ≥ 0.05 to <0.09, (5) ≥ 0.09 to <0.99, and (6) ≥ 0.99. We compared distributions of *P*-values across journals, study designs, types of analyses (primary, ordinary, baseline comparison from a randomized controlled trial (RCT), parsimonious model, interaction test, heterogeneity test, test with adjustment for multiplicity), and locations within a paper (abstract, tables, figures, text). The aim of these analyses was descriptive, and given the presence of biased sampling and lack of independence of observations, we refrained from statistical tests for these comparisons.

To obtain a more detailed distribution, we also defined 100 one-percent wide intervals of *P*-values and obtained frequencies for each interval. We plotted the logarithm of each frequency against the logarithm of the midpoint of each interval. In interpreting this distribution, we assume that the observed *P*-values come either from null hypotheses or from alternative hypotheses [15,16]. If the null hypothesis H_0 is true, and the test statistic is continuous, the distribution of *P*-values will be uniform [15,16]. This is