Reproducibility issues in science, is P value really the only answer?
Jean Gaudart, Laetitia Huiart, P. J. Milligan, Rodolphe Thiebaut, Roch Giorgi

To cite this version:
Jean Gaudart, Laetitia Huiart, P. J. Milligan, Rodolphe Thiebaut, Roch Giorgi. Reproducibility issues in science, is P value really the only answer?. Proceedings of the National Academy of Sciences of the United States of America, National Academy of Sciences, 2014, 111 (19), pp.e1934. <http://www.pnas.org/content/111/19/E1934.full>. <10.1073/pnas.1323051111>. <hal-01307492>

HAL Id: hal-01307492
https://hal-amu.archives-ouvertes.fr/hal-01307492
Submitted on 26 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License
Reproducibility issues in science, is P value really the only answer?

Johnson describes the lack of reproducibility of scientific studies, attributed, according to the author, to the low level of significance (1). We appreciate the quality of this work and its importance for the interpretation of statistical evidence. These results should be considered in statistical guidelines. Nevertheless, we would like to point out some important points not thoroughly discussed in this publication.

Not publishing “nonsignificant” results leads to the well-known publication bias whereby studies with low statistical power are underrepresented. This bias would become more severe, despite recommendations to allow for publication of “negative” results. Lowering the significance level will further increase the type II error, which is clinically as important as type I error. Focusing only on the type I error may lead to an excessive false nondiscovery rate. In the case of severe diseases, it is not uncommon to fix a significance level at 0.1 (2), at the early stages, to avoid excluding an effective treatment. Johnson argues that this may be corrected by increasing the sample size. However, increasing the size of clinical trials will reduce their feasibility and increase their duration. Aside from these issues, including more patients means exposing more patients to an experimental treatment and may challenge the equipoise concept.

The issue of fixing a threshold defining significance refers to the Fisher–Pearson controversy. Estimating a P value is needed to quantify the strength of evidence. However, fixing a threshold is needed to make a decision controlling for the risk of type I and type II error. Actually, regarding the issue addressed by Johnson, it would be interesting to assess if a priori specification of the threshold is required, or if research results could be compared using the P value and the magnitude of the tested statistic.

The issue of significance level is only the tip of the iceberg. Indeed, design issues should not be overlooked when discussing lack of reproducibility. Selection bias leads to extrapolation of results to a population different from the target population (3). Furthermore, the “poor reporting” practice highlighted by Altman et al. (4) and the lack of compliance to reporting recommendations (e.g., Consolidated Standards of Reporting Trials) hinder a proper assessment of the quality of the study and hide selection bias or misuse of statistical tests; the latter leads to nonreproducibility of the reported research. In an extreme example, monthly American Air passengers and the Australian electricity production in the late 1950s are highly correlated (Pearson’s correlation = 0.88, P = 8.8 × 10⁻¹³) without any meaning.

The causality criteria defined by Hill (5) highlight other important considerations in the interpretation of results. Reliance on P values remains surprisingly widespread, but good decision making depends on the magnitude of effects, the plausibility of scientific explanations of the mechanism, and the reproducibility of the findings by others.

Jean Gaudart,1, Laetitia Haïart,2 Paul J. Milligan,3 Rodolphe Thiebaut4, and Roch Giorgi5
1Aix-Marseille Université, Institut National de la Santé et de la Recherche Médicale, Unité Mixte de Recherche 912, 13005 Marseille, France; 2Clinical Epidemiology Department, Centre Hospitalier Universitaire La Réunion, 97405 Saint-Denis, France; 3London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom; and 4Université de Bordeaux, Institut National de la Santé et de la Recherche Médicale, Research Centre Epidemiology and Biostatistics, Institut de Santé Publique, d’Épidémiologie et de Développement, 33076 Bordeaux, France


Author contributions: J.G., L.H., P.J.M., R.T., and R.G. wrote the paper. The authors declare no conflict of interest.

To whom correspondence should be addressed. E-mail: jean.gaudart@univ-amu.fr.