CRITICAL CARE

## EDITORIAL

# Simple screening statistical tools to detect reporting bias: when should we ask for raw data?

Romain Pirracchio[1,2],*, Matthieu Resche-Rigon[2], Sylvie Chevret[2] and Didier Journois[1]

Recent accusations concerning the veracity of published data raise the question of whether clinicians should trust the results published, and how evidence-based results should be translated into clinical practice [1-3]. As editors, reviewers or readers, it is our responsibility to appraise the data published or proposed for publication, before translating the results into clinical practice. Simple screening tools for detecting certain types of reporting bias would be of interest. The evaluation may rely on the following steps: evaluation of the distribution of the reported variables; evaluation of the distribution of the reported *P*-values; parametric bootstrapping and explicit computation of the *P*-values.

In many papers, data are reported as if they were normally distributed, using summary statistics such as means and standard deviations that might not adequately reflect the distribution of non-symmetrically distributed variables [4]. One should first question whether a given variable could intrinsically behave normally or not. For instance, duration variables are usually asymmetrically distributed. Second, the reported summary statistics can provide information on the distribution. When a strictly positive variable has a standard deviation close to or even larger than its mean, the variable distribution is wide, and, if negative values are impossible, its distribution is likely to be asymmetric. Moreover, parametric statistical tests are frequently used, when they might be inappropriate if the sample size is not large enough or the distribution is too skewed. Alternatives might either rely on using non-parametric statistical tests or only on comparing the confidence intervals without any statistical tests [5].

Statistical testing should be avoided when evaluating covariate balance since randomization should produce exchangeability. While such tests are not appropriate to accept the null hypothesis, manuscripts often report such statistical tests. The distribution of the corresponding *P*-values can then be analyzed. If randomization was adequate, baseline characteristics distribution should be balanced between groups and the *P*-values referring to the comparison of baseline independent characteristics should follow a uniform distribution over the interval [0,1] [6]. In case of fraud, the authors are tented to produce *P*-values all close to 1. This should probably be considered as a warning signal. However, this relies on the fact that all baseline comparisons are reported; otherwise, an adequacy test to uniform distribution could be biased because of missing - and potentially informative - *P*-values. One could finally compute the *P*-values and compare them to the reported values.

The parametric bootstrap is a simulation procedure that consists in randomly generating many independent datasets on the basis of the reported characteristics of the population and some knowledge about the distribution law. Simulations can help to appraise two questions: is the data distribution credible and what is the probability that the authors had to find the results they report? A similar approach known as the 'detectable effect size' has been proposed by Cohen [7]. Such approaches do not aim at providing new inference but at detecting potential inconsistencies.

Simple evaluations of the reported data might offer some alert signals, although they do not enable one to conclude whether error or fraud is present. Such warning signals should prompt a request for raw data, to critically appraise the quality of the data using additional tools [8]. When dealing with potentially fraudulent data, it seems crucial to offer a multiple line screening. As stated by Haldane ('second order faking'), when data are fabricated to pass certain statistical tests, they are likely to fail on others [9].

---

*Correspondence: romain.pirracchio@egp.aphp.fr
[1]Service d'Anesthésie Réanimation, Hôpital Européen Georges Pompidou, 20, rue Leblanc, 75908 Paris, France
Full list of author information is available at the end of the article

BioMed Central

**Author details**
[1]Department of Anesthesiology and Critical Care Medicine, Hôpital Européen Georges Pompidou, Université Paris-Descartes Sorbonne Paris Cité, Paris, France. [2]Department of Informatics and Biostatistics, Unité Inserm UMR S717, Hôpital Saint Louis, Université Paris 7 Diderot, Paris, France.

**References**
1. Shafer SL: **Notice of retraction.** *Anesth Analg* 2010, **111:**1567.
2. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, Rubin HR: **Why don't physicians follow clinical practice guidelines? A framework for improvement.** *JAMA* 1999, **282:**1458-1465.
3. Grol R, Grimshaw J: **From best evidence to best practice: effective implementation of change in patients' care.** *Lancet* 2003, **362:**1225-1230.
4. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF: **Improving the quality of reporting of randomized controlled trials. The CONSORT statement.** *JAMA* 1996, **276:**637-639.
5. Faulkner C, Fidler F, Cumming G: **The value of RCT evidence depends on the quality of statistical analysis.** *Behav Res Ther* 2008, **46:**270-281.
6. Sellke T, Bayarri MJ, Berger JO: **Calibration of ρ values for testing precise null hypotheses.** *Am Stat* 2001, **55:**62-71.
7. Cohen J: *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Routledge Academic.
8. Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, Lesaffre E, Murray G, Edler L, Hutton J, Colton T, Lachenbruch P, Verma BL: **The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials.** *Stat Med* 1999, **18:**3435-3451.
9. Haldane J: **The faking of genetic results.** *Eureka* 1948, **6:**21-28.