Received: 4 July 2018 | Revised: 29 October 2018 | Accepted: 30 October 2018

DOI: 10.1002/bimj.201800195

SHORT COMMUNICATION

Biometrical Journal

A note on tests for relevant differences with extremely large sample sizes

Andrea Callegaro[1] | Cheikh Ndour[2] | Emmanuel Aris[1] | Catherine Legrand[2]

[1]GSK Vaccines, Rue de l'Institut 89, 1330 Rixensart, Belgium

[2]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Universite Catholique de Louvain, Louvain-la-Neuve, Belgium

Correspondence
Andrea Callegaro, GSK Vaccines, Rue de l'Institut 89, 1330 Rixensart, Belgium.
Email: andrea.x.callegaro@gsk.com

Funding information
BEWARE FELLOWSHIPS Industry

Abstract

A well-known problem in classical two-tailed hypothesis testing is that *P*-values go to zero when the sample size goes to infinity, irrespectively of the effect size. This pitfall can make the testing of data consisting of large sample sizes potentially unreliable. In this note, we propose to test for relevant differences to overcome this issue. We illustrate the proposed test a on real data set of about 40 million privately insured patients.

KEYWORDS

two-tailed hypothesis tests, large sample size, testing for relevant differences

1 | INTRODUCTION

In inferential statistics one of the most frequently used techniques is the hypothesis test, in particular the two-sided hypothesis test. A classical two-sided test is specified as:

$$H_0 : \Delta = 0; \quad H_1 : \Delta \neq 0,$$

where Δ is the parameter of interest.

Although this test provides valuable information on the significance of results in several cases, it is not limitation-free and must be used in the correct setting. A well-known issue with such a test is that the type II error (i.e., the probability of not rejecting H_0 when the data are actually coming from a population under H_1) is controlled only for a prespecified given alternative hypothesis $H_1 : \Delta = \delta$. If the sample size is larger than the one required, the power of the test increases dramatically. When the sample size is extremely large, a statistical test will almost always show a significant effect even if the effect size is very small and potentially meaningless. This is because the standard error (SE) shrinks as the sample size increases, producing a small *P*-value irrespective of the effect size (Lin, Lucas, & Shmueli, 2013).

This phenomenon is particularly problematic when considering large databases with millions of observations. Databases containing extremely large number of observations are nowadays routinely available, and the computer technology available today allows us to analyze such data in a (relatively) short computational time.

This issue seems to not be considered as a real problem in the medical literature and is most often solved by adding a statement about the need to differentiate clinical/biological relevance from statistical relevance, or by advising not to trust *P*-values when the sample size is too large. Lin et al. (2013) suggested to (i) report size effect and the confidence intervals; (ii) adjust the threshold *P*-value downward as the sample size grows; (iii) compute *P*-values on "small" subsets of the data.

In this note, we recommend testing for relevant differences (Cohen, 1962; Hodges & Lehmann, 1954; Victor, 1987; Wellek, 2010) as a solution to the above-mentioned problem. This test was recently brought back to attention by Wellek (2017), but it is still rarely used in practice.

TABLE 1 Incidence rate of emergency room visits due to a specific viral infection by season (subjects from Truven Health MarketScan Research Database). *N*: number of subjects; PY: number of patient years; *n*: number of ER visits; IR: incidence rate × 10,000; $\hat{\Delta}$: log incidence ratio of two consecutive years; $\hat{\sigma}_{\Delta}$: standard error of the log incidence ratio.

Years	*N*	PY	*n*	IR	$\hat{\Delta}$	$\hat{\sigma}_{\Delta}$
2005	13823952	13742748	2314	1.684	-	-
2006	20577554	20477206	3783	1.847	0.0927	0.0264
2007	22136295	21994450	5134	2.334	0.2339	0.0214
2008	30574541	30417230	6723	2.210	−0.0546	0.0185
2009	29588024	29442487	6739	2.289	0.0349	0.0172
2010	31287536	31127362	7346	2.360	0.0306	0.0169
2011	35431439	35273814	6341	1.798	−0.2722	0.0171
2012	27489436	27376983	6051	2.210	0.2066	0.0180
2013	27560720	27439647	4984	1.816	−0.1963	0.0191
2014	18566638	18505616	3665	1.980	0.0865	0.0218

2 | METHODS

The test for relevant differences is an extension of the classical two-sided approach, which tests the following hypotheses:

$$H_0 : |\Delta| \leq \delta_E; \quad H_1 : |\Delta| > \delta_E.$$

Here, $\delta_E \geq 0$ represents a small and irrelevant effect (irrelevance limits in the jargon of modern biostatistics).

Under the null hypothesis $|\Delta| = \delta_E$, the *P*-value is given by

$$P\text{-value} = \Phi\left(\frac{\delta_E - |\hat{\Delta}|}{\hat{\sigma}_{\hat{\Delta}}}\right) + \Phi\left(\frac{-\delta_E - |\hat{\Delta}|}{\hat{\sigma}_{\hat{\Delta}}}\right),$$

where $\hat{\sigma}_{\hat{\Delta}}$ is the standard error (see Chapter 11 of Wellek (2010) and Appendix). This *P*-value can be easily computed using standard software, for example using the function `pnorm` in R or the function `cdf('NORMAL',...)` in SAS.

3 | RESULTS

We illustrate the proposed test in an analysis of the incidence of emergency department (ED) visits, during which a viral infection was detected in 2013 and 2014. In this real data case study, each season yearly analysis period was defined as starting on the 1st July and ending on June 30 of the following calendar year, so that each period contained a winter season. The incidence rate (IR) was calculated by dividing the number of ED visits (n) by the number of patients (N), and multiplying this value by 10,000 (Table 1). The data used in this analysis was obtained from the Truven Health MarketScan Research Database. This substantial data source contains US individual-level healthcare claims information from large employers and managed care organisations on over 56 million patients. It complies with the Health Insurance Portability and Accountability Act of 1996 for the preservation of participant anonymity and confidentiality (Adamson, Chang, & Hansen, 2005). The IRs of selected privately insured individuals visiting an ED in the US from 2005 to 2014, are shown in Table 1, where sample sizes ranged from approximately 18.5–27.5 million.

In this case study, we addressed whether significant differences exist between the IRs of patients visiting an ED in the US between 2013 and 2014. To address our question, we used the log of the incidence rate ratio (IRR) as statistic ($\hat{\Delta} = 0.0865$ and $\hat{\sigma}_{\hat{\Delta}} = 0.0217$). Performing a classical 2-sided test (normal approximation, $Z = \hat{\Delta}/\hat{\sigma}_{\hat{\Delta}}$) produced a *P*-value of 0.00007. This small *P*-value was mainly driven by the small standard error that was present as a result of the very large sample size. In fact, the increase in incidence was relatively small (9%).

Note that, similarly to what is done when defining equivalence margin in equivalence trials, external data, evidence from scientifically sounded literature or existing international guidelines could be used to define the irrelevance limit δ_E. However, for the sake of our illustration, we consider data from the previous years to derive it. Assuming that the variability between 2005 and 2012 observed in the database represents the natural variability of the event, IRRs calculated between 2005 and 2012

inclusive (Table 1) were used to define the irrelevance limit, δ_E. A random effects meta-analysis was performed, yielding the following results: average effect = 0.0386, se = 0.064 (95% Confidence Interval [CI]: −0.0880; 0.1652) and variance of the random effect $\tau^2 = 0.0288$. Based on these results, we determined the irrelevant limit as a fraction (f = 0.5) of the upper limit of the CI: $\delta_E = 0.0826$. This is to ensure that if there is a difference between 2013 and 2014 it would at least preserve a substantial fraction of the 95% CI upper estimate of the natural variability. The *P*-value of this test was 0.42, and the null hypothesis that there is no relevant difference in the rate of patients visiting an ED between 2013 and 2014 cannot be rejected.

4 | DISCUSSION

In this note, we recommend the statistical test for relevant differences (Cohen, 1962; Hodges & Lehmann, 1954; Victor, 1987; Wellek, 2010, 2017) for data consisting of very large sample sizes. As shown in this paper, this test can be used to solve the issue of classical significance testing on very large sample sizes, providing more interpretable significance values.

It is worth noting that this approach can also be useful on high-dimensional data, such as microarray data. Such data consist of a large number of genes (p) but a small sample size. Analysis of this data type allows the identification of differentially expressed genes between the control and treated samples. In this case, p log_2 fold changes are estimated ($\hat{\Delta}_i$ and $\hat{\sigma}_{\hat{\Delta}_i}$, $i = 1, ..., p$), where there is a high chance that $\hat{\sigma}_{\hat{\Delta}_i}$ falls close to zero for certain genes, producing small *P*-values irrespective of the effect size $\hat{\Delta}_i$ (false-positive genes). This problem can also be solved by testing for relevant differences with a $\delta_E = \log_2(1.5)$, as a fold change of 1.5 is often considered as the minimum interesting/replicable effect for microarray data.

The drawback of the proposed approach with respect to the classical two-sided test ($\delta_E = 0$), is that prespecifying δ_E can be challenging in certain cases. However, approaches for determining the equivalence margins can be adapted to prespecify δ_E.

Although the statistical test for relevant differences has recently reemerged in the literature (Wellek, 2017), its practical use remains rare. Applying this approach should help researchers to obtain reliable significance testing results from datasets with extremely large samples.

ACKNOWLEDGMENTS

Medical writing services and editorial assistance and publication coordination were provided by Sonia Norris and Sophie Timmery (XPE Pharma & Science on behalf of GSK), respectively.

AUTHORS CONTRIBUTIONS

All authors were involved in the conception of the work. All authors contributed to the analysis and interpretation of the results and were involved in all stages of the manuscripts development, and approve its final version.

COMPETING INTERESTS

GlaxoSmithKline Biologicals SA was the funding source for all costs associated with this work, and the development and publishing of the present manuscript. Andrea Callegaro and Emmanuel Aris are employees of the GSK group of companies. Andrea Callegaro and Emmanuel Aris own stock options in the GSK group of companies. The institution of Cheikh Ndour and Catherine Legrand received a grant from the Walloon Region in the frame of this work.

ORCID

Andrea Callegaro https://orcid.org/0000-0003-4805-5118

REFERENCES

Adamson, D. M., Chang, S., & Hansen, L. G. (2005). Health Research Data for the Real World: The MarketScan Databases. White paper from the Research and Pharmaceutical Division of Thomson Medstat: "https://www.researchgate.net/publication/281570301"

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypothesis. *Journal of the Royal Statistical Society, Series B.*, *16*, 262–268.

Lin, M., Lucas, H. C. J., & Shmueli, G. (2013). Too big to fail: Large samples and the *P*-value problem. *Information Systems Research*, *24*, 1–12.

Victor, N. (1987). On clinically relevant differences and shifted null hypotheses. *Methods of Information in Medicine*, *26*, 155–162.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Wellek, S. (2017). A critical evaluation of the current "*p*-value controversy". *Biometrical Journal*, **59**(5), 854–872.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Callegaro A, Ndour C, Aris E, Legrand C. A note on tests for relevant differences with extremely large sample sizes. *Biometrical Journal*. 2019;61:162–165. https://doi.org/10.1002/bimj.201800195

APPENDIX

In the following, we assume that $\sigma_{\hat{\Delta}}$ is known. If it is unknown, we can replace $\sigma_{\hat{\Delta}}$ by his estimator $\hat{\sigma}_{\hat{\Delta}}$, and the Normal distribution by a Student distribution. In case of large n, the method remains unchanged, since the Student distribution, and its percentile are well approximated by a Normal distribution.

Assuming that $\hat{\Delta} \sim N(\Delta, \sigma_{\hat{\Delta}})$, the distribution of $Y = |\hat{\Delta}|$ is folded Normal with cumulative distribution function given by

$$F_Y(y; \Delta, \sigma_{\hat{\Delta}}) = \Phi\left(\frac{y - \Delta}{\sigma_{\hat{\Delta}}}\right) + \Phi\left(\frac{y + \Delta}{\sigma_{\hat{\Delta}}}\right) - 1$$

for $y \geq 0$, and 0 everywhere else.

Under the null hypothesis $|\Delta| = \delta_E$, the P-value can be obtained by

$$P\text{-value} = 1 - F_Y\left(y; \Delta = \delta_E, \sigma_{\hat{\Delta}}\right) = \Phi\left(\frac{\delta_E - y}{\sigma_{\hat{\Delta}}}\right) + \Phi\left(\frac{-\delta_E - y}{\sigma_{\hat{\Delta}}}\right),$$

where y is the observed value of the test statistic Y.

Rejecting the null hypothesis of irrelevant differences when this P-value does not exceed α is equivalent to applying the rejection rule which, according to Wellek (2010, §11.3.1), yields an uniformly most powerful unbiased test of H_0.