American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Mini Review

Between Biological Relevancy and Statistical Significance - Step for Assessment Harmonization

Tomasz Grabowski[1]*, Agnieszka Tomczyk[2], Anna Wolc[3] and Shayne Cox Gad[4]

[1]*Polpharma Biologics Trzy lipy, Gdańsk, Poland*

[2]*Parexel International, Parexel Polska Sp. z o.o., Żwirki i Wigury, Poland*

[3]*College of Agriculture and Life Sciences, Iowa State University, USA*

[4]*Department of Animal Science, Gad Consulting Services, USA*

***Corresponding author:** Tomasz Grabowski, Polpharma Biologics Trzy lipy 3, 80-172 Gdańsk, Poland ,
Email: tomasz.grabowski@polpharmabiologics.com

To Cite This Article: *Tomasz Grabowski, Agnieszka Tomczyk, Anna Wolc, Shayne Cox Gad. Between Biological Relevancy and Statistical Significance - Step for Assessment Harmonization. Am J Biomed Sci & Res. 2021 - 13(5). AJBSR.MS.ID.001908.* DOI: 10.34297/AJBSR.2021.13.001908.

Received: July 15, 2021; **Published:** July 29, 2021

Abstract

The question of whether study results are significant, relevant and meaningful is the one to be answered before every study summary and presenting conclusions. This paper analyzes and juxtaposes currently used methods to assess statistical significance, effect size, and highlights the value of understating and assessing biological relevance. Many opinions of experts in various fields are cited to demonstrate the ambiguity of merely p-value usage. The answer to the question of the best approach is complex and a 3-step approach is suggested taking into consideration

a. Statistical assessment of differences between groups

b. Effect analysis and

c. Biological relevance assessment.

The paper emphasizes the need to take into account more than just statistical significance in the decision process, or decisions on accepting or rejecting hypotheses. p-values or any other statistical tool is not recommended as the main criterion for decision making. Furthermore, none of the above mentioned 3 steps should be used in isolation to assess the results. Moreover, there is a need for publication of negative results unless directly caused by poor design or low sample size because the current tendency to focus entirely on positive results biases the literature and leads to unnecessary replication of experiment.

Keywords: Relevant, Significant, FDA, EMA, p-hacking

Minireview

The question of whether the effect of a xenobiotic on a living organism can take place usually requires a complex statistical evaluation. The question is asked from the perspective of pharmacometric analysis, toxicological studies, and analyzes of the level of impurities or xenobiotics residues in relation to human or animal safety. The final assessment related to confirmation of expected effects (hazard, curative value, toxicity, etc.) must be preceded by statistical analysis. For some studies, e.g., bioequivalence analysis, population analysis, etc. strictly defined methods of data analysis are recommended [1-3]. However, there are many settings in which guidance will and cannot be that comprehensive to explain how to interpret the clinical or biological relevance of findings beyond statistical evaluation. For most scientific studies that perform comparative analyses between groups, the ways in which data should be analyzed are not as accurately described.

It is however worth mentioning that some statistical associations have developed guidance and decision trees [4]. There is still a lot of discussion about the weaknesses in the assessment of biological effects based only on statistical analysis. So far, however,

it has not been possible to harmonize and define a unified approach to such a process [5-11] that would allow systematic assessment of relevancy of biological effects. The aim of the presented work is to propose a systematic approach to the assessment of different biological data (continuous/categorical/binary). A stepwise procedure is proposed (three-step assessment) which harmonizes guidelines and experience in various types of pharmacometric, toxicological, and other studies related to the analysis of biological effects on living organisms.

First step of assessment – differences between groups. Whether a statistical significance can be the only basis for assessing the complex, often multidimensional phenomena occurring in a living organism is currently undergoing discussion [6,9,12]. This problem has arisen, among others, as a result of "overestimating" the possibilities resulting from determining the p-value and statistical significance versus biological relevancy [13,14]. Currently, "p-hacking" or "asterisk hunting" is discussed in relation to the use of such an approach to significance analysis where initial analysis shows a lack of statistical significance but the final analysis shows significant differences between groups or an adequately powered study [15-17]. Lack of significance in underpowered studies is expected as the lower the sample size the stronger effect is required to rich significance. The influence of this problem can be remedied by publishing not only p-values but also effect sizes and standard errors. This will allow meta-analyses combining smaller studies to obtain a sufficient sample size. Meanwhile, there is an agreement that the p-value alone as a key factor in determining the significance of effect may have a very limited informative value. p-values define evidence only in relation to a single hypothesis, and therefore may not be useful when analyzing a complex biological response [5]. This is why in the case of clinical studies, sometimes a 'fragility index' is recommended for p-value verification and study robustness analysis [18]. The arguments for such an approach are cited by many authors in various fields:

a) Lack of reproducibility of high-quality studies using statistical significance for final judgment [6].

b) In some studies, in psychology p-values may not be useful [9,19].

c) p-value, or statistical significance, does not measure the effect size, moreover large and increasing sample sizes that lower p-values [8,12,19].

d) Hypothesis testing, at best, only highlights possibly interesting correlations. But such correlations almost always can be determined, no matter if they are indicated as "significant" by that measure. They are not a representation of cause [20] In observational studies correlations irrespective of size do not prove causation, acyclic directional graphs have been suggested for evaluating causal effects from observational studies [21].

e) Belief in a null hypothesis as an accurate representation of the population sampled is confronted by a logical disjunction: Either the null hypothesis is false, or the p-value has attained by chance an exceptionally low value (Fisher's Argument) [20] but at least with statistical methods error rate can be controlled by the researcher.

f) p-values are often adjusted for multiple tests using many correction factors (Bonferroni, Holm, Hochberg, Hommel) such corrections are usually not used in a consistent way [20] better guidance may be needed however the assumptions behind these correction factors are known and can be used to determine the appropriate one, also with Bayesian methods probabilities can be assigned to obtaining set effect size by a chance

g) There are more and more recommendations to shifting the p-value for example to alpha 0.005. A huge problem with predefining and setting the alpha level may be that importance of type I and II varies between studies, areas, and researchers [22]. Give all effects SE and p-values and the reader can decide which ones to follow. Making the threshold more strict will further bias the literature (Bulmer effect).

In some guidance, depending on study aim, health authorities (HA) explicitly recommend skipping statistical significance analysis. For example, biologically significant adverse effects should be used for no observed adverse effect level (NOAEL) calculations even if they are not statistically significant [23]. The stage of preliminary data analysis is currently proposed to be replaced by Bayesian analysis or determination of confidence intervals (CI) [7,24]. Another approach could be a move from pure hypothesis testing to predictive models based on predictive probability, which can be verified against real data [20]. This however requires access to preferably several independent data sets to be used for validation. Further possibilities for replacing significance testing may include equivalence tests, likelihood ratios, or information criteria [22]. Another alternative would be using power analysis to focus on sample size based on the desired width for confidence intervals or on the closeness of the sample statistics to their corresponding population parameters [22]. However, in the studies strongly focused on the 5R rule often unmeasured factors contribute to increased variation and lead to effectively underpowered study [25].

The second step of assessment–effect analysis. A typical approach in many scientific studies is to build conclusions or final evaluation of the study solely based on the statistical significance of the difference between groups. Often, apart from determining factors limiting the research model, no further assessment

elements are implemented. But the answer to the question of how to proceed to objectively assess the effect (after analysis of statistical significance) has already been described in some fields. For example, significance analysis between doses in parallel dose-response studies is not necessary if a statistically significant trend (upward slope) across doses can be determined [26]. Indeed, trend analysis is directly related to the analysis of the effect features – effect analysis. An example of the second step in biological significance/relevancy of the effect analysis in clinical trials is minimal clinically important difference calculated by different methods (distribution base; anchor-based; Delphi) [27]. One of the more widely known methods used in this case is represented by effect size calculation which helps describe the magnitude of differences [7,10,28]. Depending on the nature of the analysis of effect size using methods like Odds ratio (OR), Cohen's d, Cohen's f2 Hedge's g, Glass Δ and Δ*, Steiger's Ψ, Pearson's r, Spearman's ρ, Cramer's V, Chi-square ϕ, r2, adjusted r2, n-way ANOVA f2, 1-way ANOVA η2, n-way ANOVA partial η2, 1-way/n-way ANOVA ω2 are recommended [29].

Statistically significant effects or changes may not be meaningful for the general state of the system and this is why "absence of evidence is not evidence of absence" [30]. A study may be inadequately powered, e.g., due to model limitations or when the mechanism or mode of action is not completely understood or is unknown. Furthermore. e.g. biomarkers used for statistical significance analyses could be only partially linked to the biological effect (toxicological, clinical, etc.) and may not be representative of the "true effect" but are only partially estimates thereof [31]. In such cases, researchers cannot be sure which biomarker is directly and fully linked only to the desired effect. Moreover, one or two markers never can represent all interactions in the entire biological system of an animal or human. A biological effect is the representation of a continuum of changes after a drug or any other xenobiotic dosing. Quantitative indices or markers cannot represent multidimensional characteristics of that continuum so they could be only an approximation of the "true effect" [30].

In such cases, a statistical analysis based on such biomarkers could not be fully but only partially linked to the biological relationships? Even if biomarkers related to the mode of action are very sensitive, they might not be directly linked only to the expected biological effect. This problem has been described, for example, in relation to carcinogenicity studies background genes or local tissue processes [32]. It was emphasized that "even when an hypothesized mode of action is supported for a described response in a specific tissue, it may not explain other tumor responses" [32]. Then, after statistical analysis, the difference between groups could be statistically insignificant but for the desired effect, a certain relevance may, nevertheless, be indicated e.g., by effect size parameters like Cohen's d or other effect size indices. Some criteria for this step of assessment were proposed but still need evaluation: for example, the biologically relevant effect for at least 10% change in body weight in toxicology [13]; clinically relevant effects in population modelling are considered "clinically relevant" at> 20% [33]; Cohen's d > 0.8.

The third step of assessment – biological relevance assessment. Both 'significance CI analysis' and 'effect size' are elements of statistical consideration and may not always be the basis for a final assessment of the impact of a particular factor on the biological effect. This became the reason why "biologically significant effect" was defined [34]. The term "biologically significant" is defined in order to distinguish from "statistically significant" and to be used as a key element of assessment when the term "statistically significant" does not allow an adequate verification of the study results. At present, this concept has not been fully harmonized, and many terms are currently used: biologically relevant [13,35,36], biologically significant [34,36], biologically or toxicologically meaningful [35,37], noteworthy [30] biological importance [6,38], biologically unrealistic [32], clinically meaningful [26], etc.

Biological relevancy of the findings should be a separate step of analysis related to physiology and should be a matter of a mechanistic biological/pharmacological/toxicological approach. This kind of approach is part of the definition of biological significance (… to distinguish from "statistically significant"). In the case of carcinogen risk assessment studies and regulations, it was emphasized that statistically significant differences may or may not be biologically relevant [32]. The justification for this approach and the separation between statistical analysis and interpretation of the analyzed phenomenon is confirmed by some guidelines. In its guidance related to chronic toxicity and carcinogenicity studies, the Organisation for Economic Co-operation and Development (OECD) refers to "statistical analysis being a part of the interpretation of the biological importance, not an alternative" [38]. However, such determination also needs to consider sufficient sample size (in small samples large differences can be obtained due to natural variability or unmeasured stratification). Another important factor would be a determination of the outliers – their origin from biological phenomenon can be biologically significant however a risk of measurement error must also be considered.

Some good examples of the stepwise process of assessment are NOAEL calculations or toxicological relevancy assessments based on biomarkers levels [23]. In toxicology assessment, biologically meaningful changes relating to a change in biomarker levels could be concluded when there is confirmation in histopathological changes [37]. Studies conducted to investigate the effect of xenobiotics are not validated. This is why reproducibility of the findings might be challenging and further considerations based upon historical data may be needed to derive biological relevancy. Lack of biological

relevancy could be stated also if weak, equivocal, or not reproducible responses or small statistically significant differences but within CI of historical data are observed [35]. Statistics represent a valuable and essential tool in toxicology, but they are often subject to misuse. The most common form of misuse is confusing the results of the analysis with data or proof of an association between treatment and observed effect. Central here is that correlation is not proof of causality [39]. In interpreting results of a study or assessment of an endpoint (such as a drug causing liver damage), one needs to consider not just a finding of statistical significance in the difference between a control and treated group, but all the other available pieces of data that are available as well [40]. Examples of questions raising concerns are: Is statistical significance found at only one dose level? Or is there a dose-response at least across several significant dose levels? Is the finding supported by other sets of data? Are the several (potentially) associated aspects of clinical chemistry, organ weights, histopathology, and other indicators of adverse effects in alignment? Is the effect sufficient to indicate a biologically or clinically significant effect? An example here would be Hy's law that increases in liver enzymes in patients need to be 3-fold greater to be considered clinically significant [41,42].

A finding of the adverse effect that stands on statistical significance alone is weak and questionable in merit. There are other frequent errors in the use of statistics. Briefly, combining or pooling of data from nonidentical studies to achieve significance. The case of claiming that benediction was teratogenic is an example here. While there was not a single study supporting such a conclusion by plaintiffs, lawyers, or experts when numbers of multiple structural defects from multiple studies were combined, a statistically significant result could be calculated.

This was overwhelmingly rejected by experts and by the supreme court of the United States using a bidirectional and two-sided hypothesis test to evaluate the statistical significance of a single-sided hypothesis (and in toxicology, most hypothesis are single sided – did the treatment in question increase a clinical chemistry parameter or increases in tumour numbers). Using a two-sided hypothesis test in such cases serves to double the plausibility of finding a statistically sufficient outcome. Ignoring variance inflation considering lack of statistical significance as proof of lack of effect. Toxicology studies are performed with small groups of animals. In the dose-response region surrounding a threshold of effect, it is common for some animals to respond earlier/more robustly than others. This can serve to inflate the variance in a sample statistic, which in turn can preclude an effect being found to be statistically significant. One should always pay attention to measures of within-group variability when analyzing the results of a study. If the standard deviation (or error) increases greatly in a group, the meaning should be considered in combination with other available data [43].

Conclusions

Analysis of in vivo effects in pharmacology or toxicology could be at most selective. Because of the complex nature of the measured effects, they never have a chance to be fully specific. Full validation of biological models and studies related to biological effects in vivo is not possible. This is the main reason why the assessment of such studies cannot be done in a purely quantitative manner and why three-stage evaluation procedures apply in this case. The three-step approach to analyzing results in pharmacokinetic, pharmacodynamic, or toxicological studies is already to some extent described by the guidelines of various agencies or HA. Unfortunately, such an approach is not harmonized in any way. Suggestions on how to proceed are described in an arbitrary and heterogeneous manner across various documents. However, based on the experience of many different fields related to pharmacometrics, toxicology, clinical studies or drug residue analysis, etc. a structured three-step approach to assessing biological data can be proposed.

A stepwise harmonized approach to assessing the result of data analysis illustrating processes in biology seems to be the optimal way to proceed in scientific research. Significance or CI analysis as the first step, a measure of effect size as an irrespective second step. The third key assessment step should cover translation of the two previous steps of analysis to physiology, risk assessment, or clinical practice depending on study nature. The current paper shows that a stepwise procedure in biological effects assessments could be used for data analysis to help in a planned way move from statistical evaluation to conclusions.

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

This study was funded by Narodowe Centrum Badań i Rozwoju (grant number: POIR.01.01.01-00-0649/16)

Conflicts of interest/Competing interests

Author: Tomasz Grabowski, Agnieszka Tomczyk, Anna Wolc, and Shayne Cox Gad declare that they have no conflict of interest.

Availability Of Data and Material

Not applicable

Code Availability

Not applicable

Authors' Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Tomasz Grabowski, Agnieszka Tomczyk, Anna Wolc, and Shayne

Cox Gad. The first draft of the manuscript was written by Tomasz Grabowski and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Additional Declarations for Articles in Life Science Journals That Report The Results of Studies Involving Humans and/or Animals

Not applicable

Ethics Approval

Not applicable

Consent to Participate

Not applicable

Consent for publication

Not applicable

Acknowledgements

Authors would like to express great appreciation to Neil Johnson PhD for his professional guidance and valuable support in manuscript preparation.

References

1. (2014) FDA Guidance for Industry Bioavailability and Bioequivalence Studies Submitted in NDAs or INDs-General Considerations. 1-29.
2. (2010) EMA Guideline on the Investigation of Bioequivalence. 1-27.
3. (2019) FDA Population Pharmacokinetics Guidance for Industry. 1-26.
4. Kobayashi K, Pillai KS, Michael M, Cherian KM, Ono A, et al. (2014) Transition of JapaŽ s statistical tools by decision tree for quantitative data obtained from the general repeated dose administration toxicity studies in rodents. International Journal of Basic and Applied Sciences 3: 507-520.
5. Lytsy P (2018) P in the right place: Revisiting the evidential value of P-values. J Evid Based Med 11(4): 288-291.
6. Lovell DP (2013) Biological importance and statistical significance. Journal of agricultural and food chemistry 61(35): 8340-8348.
7. Lee DK (2016) Alternatives to P value: confidence interval and effect size. Korean J Anesthesiol 69(6): 555-562.
8. Tanha K, Mohammadi N, Janani L (2017) P-value: What is and what is not. Med J Islam Repub Iran 31: 65.
9. Hubbard R, Lindsay RM (2008) Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. Theory & Psychology 18: 69-88.
10. Friese M, Frankenbach J (2019) p-Hacking and publication bias interact to distort meta-analytic effect size estimates. Psychol Methods 25(4): 456-471.
11. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, et al. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology 31(4): 337-350.
12. Wasserstein RL, Lazar NA (2016) The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70(2): 129-33.
13. Committee ES, Hardy A, Benford D, Halldorsson T, Jeger MJ, et al. (2017) Guidance on the assessment of the biological relevance of data in scientific assessments. EFSA Journal 15: e04970.
14. EFSA (2011) Statical Significance and Biological Relevance. EFSA Journal 9: 1-17.
15. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD, et al. (2015) The extent and consequences of p-hacking in science. PLOS Biol 13(3): e1002106-e.
16. Chuard PJC, Vrtilek M, Head ML, Jennions MD (2019) Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? PLoS Biol 17(1): e3000127.
17. Lenth RV (2001) Some Practical Guidelines for Effective Sample Size Determination. The American Statistician 55(3): 187-193.
18. Ohl A, Schelly D (2017) Beyond p-values: A case for clinical relevance. British Journal of Occupational Therapy 80(12): 752-755.
19. Hyland P, Shevlin M, Kerig PK (2019) Journal of Traumatic Stress p Value Guidelines. Journal of traumatic stress 32(5): 651-652.
20. Briggs WM (2019) editor Everything Wrong with P-Values Under One Roof. Beyond Traditional Probabilistic Methods in Economics Cham. Springer International Publishing.
21. Bello NM, Ferreira VC, Gianola D, Rosa GJM (2018) Conceptual framework for investigating causal effects from observational data in livestock1. Journal of Animal Science 96(10): 4045-4062.
22. Trafimow D, Amrhein V, Areshenkoff CN, Barrera-Causil CJ, Beh EJ, et al. (2018) Manipulating the Alpha Level Cannot Cure Significance Testing. Frontiers in Psychology 9: 699.
23. (2005) FDA Guidance for Industry Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers 1-30.
24. Swersey AJ, Colberg J, Evans R, Kattan MW, Ledolter J, et al. (2019) Decision models for distinguishing between clinically insignificant and significant tumors in prostate cancer biopsies: an application of Bayes' Theorem to reduce costs and improve outcomes. Health Care Manag Sci 23(1): 102-116.
25. Kumar AHS (2012) Effectively communicating the 5R's (replace, reduce, refine, reuse, and rehabilitate) of research ethics, biomedical waste, personalized medicines and the rest. J Nat Sci Biol Med 1-2.
26. ICH (1994) ICH Topic E4 Dose Response Information to Support Drug Registration. European Medicines Agency science Medicines health Pp.1-10.
27. Fleischmann M, Vaughan B (2019) Commentary: Statistical significance and clinical significance - A call to consider patient reported outcome measures, effect size, confidence interval and minimal clinically important difference (MCID). Journal of Bodywork and Movement Therapies 23(4): 690-694.
28. Carneiro CFD, Moulin TC, Macleod MR, Amaral OB (2018) Effect size and statistical power in the rodent fear conditioning literature - A systematic review. PloS one 13(4): e0196258-e.
29. Ialongo C (2016) Understanding the effect size and its measures. Biochem Med (Zagreb) 26: 150-163.
30. ECETOC (200) Recognition of, and differentiation between, adverse and non-adverse effects in toxicology studies 1-50.
31. Lenth RV (2007) Post Hoc Power: Tables and Commentary. The University of Iowa Department of Statistics and Actuarial Science 378: 1-13.

32. EPA (2005) Guidelines for Carcinogen Risk Assessment 1-166.

33. Xu XS, Yuan M, Zhu H, Yang Y, Wang H, et al. (2018) Full covariate modelling approach in population pharmacokinetics: understanding the underlying hypothesis tests and implications of multiplicity. Br J Clin Pharmacol 84(7): 1525-1534.

34. Lewis RW, Billington R, Debryune E, Gamer A, Lang B, et al. (2002) Recognition of adverse and nonadverse effects in toxicity studies. Toxicol Pathol 30: 66-74.

35. ICH (2012) ICH guideline S2 (R1) on genotoxicity testing and data interpretation for pharmaceuticals intended for human use. Pp. 1-28.

36. FDA (2002) Guidance for Industry Immunotoxicology Evaluation of Investigational New Drugs pp. 1-38.

37. FDA (2016) Considerations for Use of Histopathology and Its Associated Methodologies to Support Biomarker Qualification Guidance for Industry pp. 1-18.

38. OECD (2010) OECD Guidance Document for The Design and Conduct of Chronic Toxicity And Carcinogenicity Studies, Supporting Tg 451, 452 and 453 pp. 2-44.

39. Gad SC, Weil CS (1988) Statistics and Experimental Design for Toxicologists. Caldwell NJ, et al. (Eds.) The Telford Press (2nd edn), USA.

40. Zimmerman HJ (1999) Hepatotoxicity: The adverse effects of drugs and other chemicals on the liver. Lippincott Williams & Wilkins. London.

41. Robles-Diaz M, Lucena MI, Kaplowitz N, Stephens C, Medina-Cáliz I, et al. (2014) Use of Hy's law and a new composite algorithm to predict acute liver failure in patients with drug-induced liver injury. Gastroenterology 147(1): 109-18.e5.

42. Fontana RJ, Hayashi PH, Gu J, Reddy KR, Barnhart H, et al. (2014) Idiosyncratic drug-induced liver injury is associated with substantial morbidity and mortality within 6 months from onset. Gastroenterology 147(1): 96-108.e4.

43. Rousseaux CG, Shockley K, Gad SC (2020) Experimental Design and Statistical Analysis for Toxicological Pathologists. In: Haschek WM, et al. (Eds.) Haschek and Rousseaux's handbook of toxicologic pathology (3rd edn) In press: Academic Press, USA.