Bionatura

Latin American Journal of Biotechnology and Life Sciences

Publications from the Humboldt Kolleg Ecuador 2019
"Breaking Paradigms: Towards a Multi-, Inter- and Transdisciplinary Science"
In commemoration of the 250th Anniversary of
February 21th – 24th, 2019.
Ibarra, Ecuador

250 Humboldt

Alexander von Humboldt
Stiftung/Foundation

CS 2019.02.01.3

Humboldt Kolleg Ecuador 2019
February 22th – 24th, 2019
Ibarra, Ecuador

Bionatura Conference Series Vol 2. No 1. 2019
"Breaking Paradigms: Towards a Multi-, Inter- and Transdisciplinary Science" In commemoration of the 250th Anniversary of Alexander von Humboldt

INVESTIGATION / RESEARCH

Basic antidepressant research: a brief assay on how to justify your alpha

Cilene Lino de Oliveira

Disponible en: http://dx.doi.org/10.21931/RB/CS/2019.02.01.3

ABSTRACT

Antidepressant research seems under risk of bias and poor reproducibility. Recent debates brought the use of the p values in hypothesis testing to the center of a reproducibility crisis. In basic biomedicine, the use of p values has been justified by tradition instead of reasoning. Here, a biomedical researcher commented concerns with the traditional use of the p values in basic antidepressant research and discussed the missing pieces limiting the plausible justifications to their use in the field.

Keywords: experimental design; statistical methods; biostatistics; biomedicine; animal models, preclinical studies.

INTRODUCTION

Basic research employs an experimental model that might be useful to discover new compounds with therapeutic value. In antidepressant research, experimental models are challenging and imperfect due to the complexity of mechanisms underlying the subjective symptoms of Major Depression or other affective disorders [1]. Ideally, a model for antidepressant research should recreate in laboratory symptoms and the neurobiological disturbances similar to those found in patients [2]. Additionally, the modeled symptoms and disturbances should be reversible by treatments effective in therapeutics. In other words, experimental models may be used to gain information on the potential utility of an unknown intervention in the treatment of diseases. Despite imperfections and partial validity, *in vivo* models are currently more representative than *in silico* or *in vitro* approaches in biological psychiatry [3].

The unstable nature of the biological outcomes difficult the standardization and reproducibility of *in vivo* models in

the laboratory. For example, natural or pathological features of an organism are species-age-sex-and so non-specific and, furthermore, interact with the environment where animals or humans are living. [4,5,6]. Therefore, *in vivo* settings require systematic approaches to prepare laboratory and experimental conditions to achieve consistency and internal validity [7,8]. Despite the efforts of the scientific community to increase the validity of *in vivo* studies [9,10], biological psychiatry still seems susceptible to reproducibility problems [11].

Some researchers claim that reproducibility crisis in biomedicine may be more related to statistical misuse and poor analytical decisions than to other technical aspects of the biomedicine [12]. Recent debates brought the use of the p values in hypothesis testing to the epicenter of a reproducibility crisis in sciences [13,14]. Biomedical studies are especially under scrutiny because the uses and misuses of p-values seem endemic in basic, preclinical or clinical levels [15,16]. Although no consensus has been met, many researchers advocate for the banning of the traditional use of p values from biomedicine [15,16,17,18,19]. I share with others the view that before abrupt decisions, scientists should consider the pros and cons and justify their choices [14].

In the field of basic antidepressant research, the uses of p values for hypothesis testing are on the grounds of routine and tradition instead of elaborated reasoning [20]. In this text, the emphasis is on how a biomedical researcher could interpret properly and justify the use of p values for hypothesis testing in a scientific project [21,22,23,24]. Because the range of research questions in biomedicine is vast, justifications on analytical choices may be also vast [25]. Hence, the examples to make the misconceptions and concepts clearer were taken from basic biomedical studies on the author's discretion focusing on animal models for antidepressant research. To understand the uses of p values in hypothesis testing it seems advisable to examine their definitions, history, and examples (further readings [20,26,27,28,29]. Hereafter, in the following text, I attempted to summarize general information on p values before discussing their specific applications in the field of antidepressant research.

An official definition by ASA describes p values as "*the probability under a specified statistical model that a statistical summary of data (for example, the sample means the difference between two compared groups) would be equal to or more extreme than its observed value*" [20]. The previous sentence could also be read as "*the probability (p) of the observed data with certain features, or more extreme than the observed ones occurred given they were drawn from a hypothetical population with the certain features*". The features of the hypothetical population are called parameters (mean, standard deviation[1] , etc.), while the features of the sample are called statistics (means, standard error, etc). *P values* are the probabilities associated with statistics and the higher a *p-value,* the higher a probability that the sample was *a part taken from* the hypothetical population.

In traditional hypothesis testing, the null hypothesis or H0, i.e., a population with mean equals zero and variance equals one, is the "hypothetical population with the certain features". Thus, the higher a p-value, the higher a probability that the sample was a part taken from the null hypothesis. In contrast, the lower a p-value, the lower a probability that the sample was a part taken from the null hypothesis. P-values as fiducial inference against a null hypothesis were created by Fisher and was remodeled by the frequentist views of Neyman and Pearson at the beginning of the twentieth century when the first controversy on the matter also appeared [26]. Although classical and frequentist interpretations of p values differ, both schools agree on their value as an approach to hypothesis testing [2628]. Over time, different research fields accepted p values for hypothesis testing with more or fewer deliberations [15,18,30,31,32]. Every project or research question will have a particular *null hypothesis* allowing particular conclusions depending on the probability of the sample be drawn from the *null* population.

In the context of basic biomedical research, the *null* hypotheses often may be declared as the absence of an effect or a *null* effect. Consider, for example, an idealized experiment for evaluation of a putative new antidepressant (drug A). In this experimental setting, biological outcomes (behavior, blood pressure, glycemia, etc) will be registered in subjects (animals, cells, tissues, etc) randomly assigned to different groups (e.g. control group treated with water, or experimental group treated with drug A). The values of the biological measures or outcomes or dependent variables of each group will be summarized into statistics according to their nature (e.g. paired or independent, quantitative or

qualitative, normal or non-normal, etc.). The statistical summaries (means, variances, etc.) of the outcome data allow comparison between the groups using statistical approaches. In the present example, two-sample *Student's T-test* might be a suitable approach once a *pair* of independent groups will be compared (see Box 1 for further information). In the *Student's T-test,* the *null* hypothesis may be stated as *the null difference* between the two means. Thus, the higher the *p-value* associated with the calculation of the *t value,* the higher the probability of the *null difference*. Conversely, the lower the *p-value* associated with the calculation of the *t value,* the lower a probability of the *null difference*. Conclusions of the theoretical study described above may then vary from "drug A has an effect" to "drug A has non-effect"[2] depending on the *p-value* associated with the statistical test.

In basic antidepressant research, the use of *p values* associated with statistical tests to claim "statistical significance" of scientific data seems very common. Terms such as "very significant" or "significant" or "non-significant" are traditionally used in basic biomedicine according to *p values.* Although the classification of scientific results according to their significance or importance is beneficial to the appraisal of scientific evidence, *p values* seem inappropriate for it [18]. The "significant"-related terms may reflect, at best, researchers' degree of confidence on the *null* hypothesis based on the data, without any connotation of biological value or confidence on the alternative hypothesis [19,20]. In other words, low values of *p* are evidence against *null*, not in favor of a specific alternative hypothesis, as commonly stated [19,20,21]. American Statisticians Association recently published a collection of papers advising scientists to move beyond *p values* when doing appraisals of their data [33]. So, why bother to calculate *p values* for hypothesis testing? Because *p values* may help to assess the rate of errors in hypothesis testing or assist decisions on acceptable levels of errors in experiment [22], for example.

In a frequentist view, the low values of *p* associated with a statistical test denote low probabilities of the Type I error [22,24,28]. In hypothesis testing, the Type I error means the probability, named *alpha*, of "rejecting the *H0* when it is true" (i.e., a false positive result) and Type II error represents the probability, named *beta*, of "non-rejecting *H0* when it is false" (i.e., a false negative result). In this context, *alpha* would represent the upper limit of the Type 1 error or false positive results tolerated in the experimental situation in the long run [22,24]. For example, the traditional values of *alpha* such as 0.05 or 0.01 indicate a rate of a Type I error or a false positive result occurring at every 20 or 100 replications of the experiment, respectively, everything else being equal [22,24]. Then, a*lpha* is a theoretical, arbitrary, "special" *p-value* that should be set during experimental design, i.e., before the collection of data, to control the rate of Type I error in an experimental setting or research field [22,23,24]. Depending on the research field, missing a real effect every 20 or 100 replications worth the risk while in other fields, the price to pay for this mistake may be too high, demanding the lowering of the acceptable value of *alpha*.

Benjamin et al. (2017) [13] proposed that sciences should adopt a default value of *alpha* equals 0.005, instead of 0.05, in hypothesis testing to reduce the rate of Type I error improving reproducibility. Then, when a *p-value* associated with a statistical test is lower than a low *alpha*, it indicates minimal rates of errors in the experimental setting? The answer is: no, not automatically, because setting the *alpha* value may help to control the rate of Type I error in the experimental setting *without affecting* the rate of Type II error. Moreover, the probability of Type I error is balanced by the probability of Type II error then by lowering the *alpha,* other things being equal, *beta* will increase. High rates

of Type II error may also bring inconsistent results over time contributing to reproducibility problems. Therefore, the focus on monitoring of *alpha*, without the appraisal of other features such as experimental design and statistical power (power= 1- *beta*), will do little for scientific reproducibility [14]. It is up to researchers in a research field to decide how tolerable the amounts of Type I and II errors are in an experimental research plan.

Beyond the focus on *alpha* values, some authors have discussed the suitable conditions to keep low the rate of errors in experimental settings [22,23,24]. Benning (2018) [23] provided putative justifications to analytic choices by examining experimental scenarios originated from different levels of theoretical backgrounds (exploratory or confirmatory studies) or availability of samples (abundant or scarce resources). Lakens (2018) [24] discussed an approach to reducing *alpha*, controlling *beta*, as a function of the sample size. Mudge et al. (2012) [22] performed an extensive study on the consequences of variations of *alpha* values on the amount and the balance of Type I and Type II errors in experimental settings. In this last reference, authors suggested an approach to classify results as significant in studies with low power and low sample size, which are typical in basic biomedical research (see Box 2 for further information).

CONCLUSION

What would then be a suitable justification for *alpha* levels in basic antidepressant research? I do not see a definitive answer for this last question because many pieces of information required to the analytical justifications are still missing: 01- What is the minimal effect size of interest in the field; 02- What is the suitable statistical power necessary to estimate the minimal effect size of interest in the field?; 03- What are the acceptable rates of errors Type I and II in the field? In the specific case of *in vivo* models, there is yet an extra unknown: 04- what is the ethical cost associated to the different Types of errors? Some efforts are in progress to address issues related to questions 01 and 02 [34, 35] while the aspects related to questions 03 and 04 still requires more attention and discussion in the research field.

ACKNOWLEDGMENTS

I thank the organizers of the Humboldt Kolleg “Breaking Paradigms: Towards a Multi-, Inter- and Transdisciplinary Science" in commemoration of the 250th Anniversary of Alexander von Humboldt, Ibarra, Ecuador. I am thankful to the students in my research group for discussions on the subject and also to Dr. José Marino-Neto for the critical reading of the manuscript.

REFERENCES

1. Maximino, C., & van der Staay, F. J. (2019). Behavioral models in psychopathology: epistemic and semantic considerations. *Behavioral and Brain Functions*, *15*(1), 1. https://doi.org/10.1186/s12993-019-0152-4

2. Willner, P., & Mitchell, P. J. (2002). The validity of animal models of depression. *Psychopharmacology (Berl)*. https://doi.org/10.1007/BF00427414

3. Belzung, C., & Lemoine, M. (2011). Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. *Biology of Mood & Anxiety Disorders*. https://doi.org/10.1186/2045-5380-1-9

4. Nemeth, C.L., Harrell, C.S., Beck, K. D. et al. (2013). No Title. *Not All Depression Is Created Equal: Sex Interacts with Disease to Precipitate Depression.*, *4*(8). Retrieved from https://doi.org/10.1186/2042-6410-4-8.

5. Miller, G. W., & Jones, D. P. (2013). The nature of nurture: refining the definition of the exposome. *Toxicological Sciences : An Official Journal of the Society of Toxicology*, *137*(1), 1–2.

6. Cole, J. H., Marioni, R. E., Harris, S. E., & Deary, I. J. (2018). Brain age and other bodily “ages”: implications for neuropsychiatry. *Molecular Psychiatry*, *24*(2), 266–281.

7. Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., ... & Sterne, J. A. (2011). The Cochrane Collaboration’s tool for assessing risk of bias in randomized trials. *Bmj*, d5928.

8. Hooijmans, C. R., Rovers, M. M., de Vries, R. B., Leenaars, M., Ritskes-Hoitinga, M., & Langendam, M. W. (2014). SYRCLE’s risk of bias tool for animal studies. *BMC Medical Research Methodology*, *14*(1), 4.

9. Kilkenny, C., Browne, W., Cuthill, I. C., Emerson, M., Altman, D. G., N. R. G. W. G. (2010). (160AD). Animal research: reporting in vivo experiments: the ARRIVE guidelines. *British Journal of Pharmacology*, *7*(1577–9).

10. Rooney, A. A., Cooper, G. S., Jahnke, G. D., Lam, J., Morgan, R. L., Boyles, A. L., Ratcliffe, J. M., Kraft, A. D., Schünemann, H. J., Schwingl, P., Walker, T. D., Thayer, K. A., … Lunn, R. M. (2016). How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards. *Environment International*, *617*(29), 92–93.

11. Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797.

12. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

13. Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

14. Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0311-x

15. Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. *JAMA*, *315*(11), 1141. https://doi.org/10.1001/jama.2016.1952

16. Kyriacou, D. N. (2016). The Enduring Evolution of the P Value. *JAMA*, *315*(11), 1113. https://doi.org/10.1001/jama.2016.2152

17. Ponsonby, A. L., & Dwyer, T. (2014). Statistics: Biomedicine must look beyond P values. *Nature*, *507*(7491), 169.

18. Trafmow, D., and Marks, M. (2015). “Editorial.” *Basic and Applied Social. Psychology*, *37*, 1–2.

19. McCarren, M., Hampp, C., Gerhard, T., & Mehta, S. (2017). Recommendations on the use and nonuse of the p value in biomedical research. *American Journal of Health-System Pharmacy*, *74*(16), 1262–1266

20. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

21. Panagiotakos, D. B. (2008). The value of p-value in biomedical research. *The Open Cardiovascular Medicine Journal*, *2*(97).

22. Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PloS One*, *7*(2), e32734.

23. Benning, S. D. (2018). How to justify your alpha: step by step. Blog Psychophysiology of Emotion and Personality Laboratory. Retrieved from http://www.peplab.org/

24. Lakens, D. (2018). Justify Your Alpha by Decreasing Alpha Levels as a Function of the Sample Size. The 20% Statistician. Retrieved from http://daniellakens.blogspot.com/

25. Flier, J. S., & Loscalzo, J. (2017). Categorizing biomedical research: the basics of translation. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *31*(8), 3210–3215.

26. Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*(1), 1–32.

27. Bayarri, M. J., & Berger, J. O. (2004). he interplay of Bayesian and frequentist analysis. *Statistical Science*, 58–80.

28. Mayo D. G. & Spanos A. Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. Brit. J. Phil. Sci. 57 (2006), 323–357.

29. Vallverdú, J. (2015). *Bayesians versus frequentists: A philosophical debate on statistical reasoning.*

30. Nuzzo, R. (2014). Scientific method: statistical errors. *Nature*, *506*(7487), 150.

31. Boyd, J. C., & Annesley, T. M. (2014). To P or Not to P: That Is the Question. *Clinical Chemistry*, *60*(7), 909–910. https://doi.org/10.1373/clinchem.2014.226282

32. Grabowski, B. (2016). "P < 0.05" Might Not Mean What You Think: American Statistical Association Clarifies P ValuesNo Title. *Journal of the National Cancer Institute*, *108*(8), 1.

33. Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p< 0.05". *The American Statistician*, *73*(Editorial).

34. Kara, N. Z., Stukalin, Y., & Einat, H. (2018). Revisiting the validity of the mouse forced swim test: Systematic review and meta-analysis of the effects of prototypic antidepressants. *Neuroscience & Biobehavioral Reviews*, *84*, 1–11.

35. Ramos-Hryb, AB, Harris, C., Aighewi, O., Lino-de-Oliveira, C. (2018). How would publication bias distort the estimated effect size of prototypic antidepressants in the forced swim test? *Neuroscience & Biobehavioral Reviews*, *92*, 192–194.

Received: 1 May 2019
Approved: 3 June 2019

Prof. Dr. Cilene Lino de Oliveira (Ph.D., Associate Professor in Physiology and Pharmacology). Department of Physiological Sciences, Biological Sciences Center, Federal University of Santa Catarina.

Correspondence should be addressed to: Dr. Cilene Lino de Oliveira: Departamento de Ciências Fisiológicas, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Campus Universitário Trindade, 88049-900 – Florianópolis – SC – Brazil. Phone: +55 48 3721 7085, Fax: +55 48 3721 9672, E-mail: cilene.lino@ ufsc.br