Power and True Report Probability in the Literature on Mice Nicotine Conditioned Place Preference

Léonard F.[1], Tirelli E.[1]

[1] University of Liège

Background

A lack of prospective power and use of effect sizes in the literature of various fields have been revealed and characterized over the years, giving rise to serious doubts on the reproducibility of many scientific results (Button et al., 2013; Cohen, 1962). To our knowledge, no study has address this problem in the field of experimental psychopharmacology using animal models.

Objective

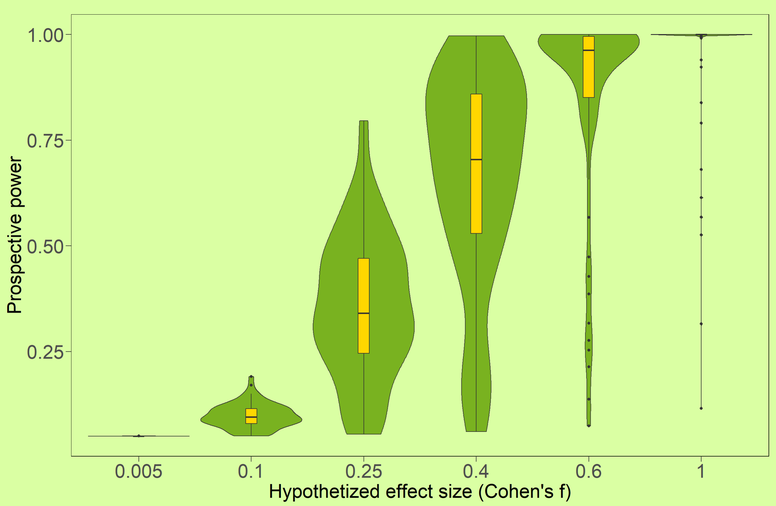
Underpowered studies present low true report probabilities (TRP). TRP is the (posterior) probability that the alternative hypothesis is true (that the effect exists) if a significant result is found (Szucs & Ioannidis, 2017). It is also named Positive Predictive Value (PPV) and requires the use of prior probabilities (plausibility). This study aimed at determining the true report probability (TRP) of representative articles examining nicotine-induced conditioned place preference (CPP) in mice. CPP is commonly used in drug addiction experimental psychopharmacology.

Methods

The articles were identified in PubMed. The sample size, the type of statistical test, its result, degrees of freedom and p-value were recorded. We then computed the individual and the median prospective powers for 6 possible effect sizes (Cohen's f: 0.005, 0.1, 0.25, 0.4, 0.6, 1; d/2). The TRP was computed from the median power ($1-\beta$), type-I error rate (α) and the plausibility (prior probability, R) of the effect.

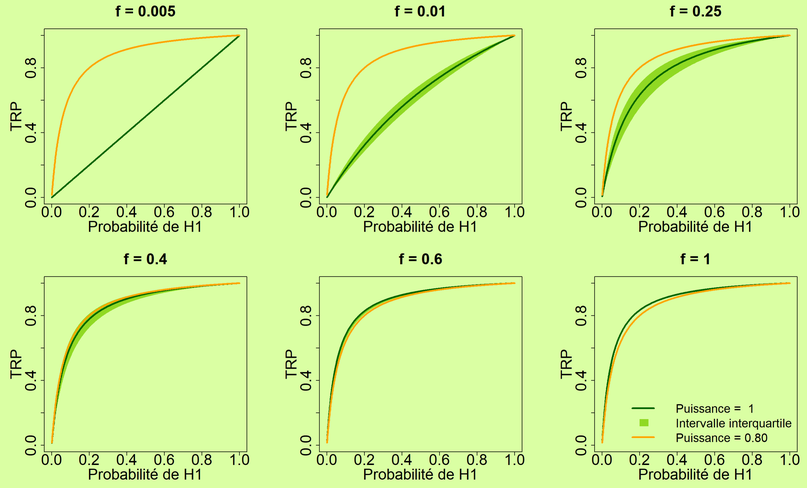
Results

Figure 1 : Prospective powers



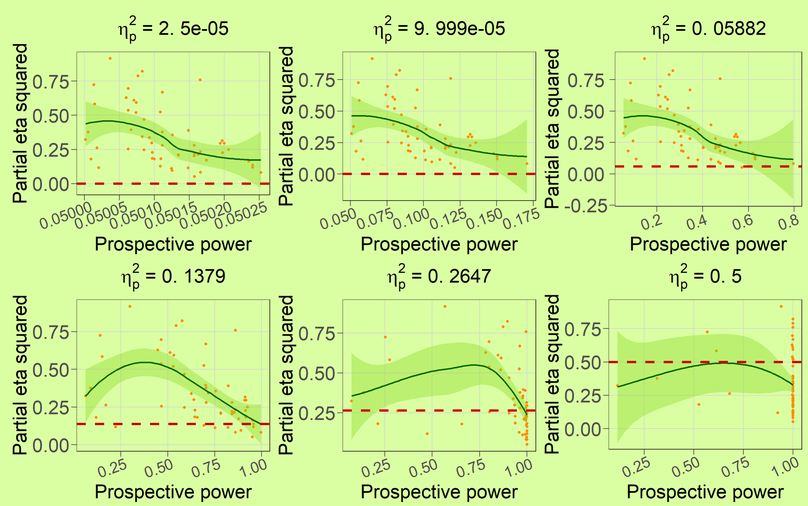
Prospective powers computed for 6 hypothetical effect sizes and 70 extracted F tests. Within each violin, a boxplot represents the median, the interquartile range (IQR) and minimum and maximum (min = Q1-(1.5IQR), max = Q1+(1.5IQR)). Points represent power outside of the boxplot and the dashed lines represent the recommended power of 80%. Amongst 139 identified articles, 47 met our inclusion criteria (esp, complete statistical results) for 70 F tests. In this sample 81% of tests were significant. The median power for small (0.2), medium (0.5) and large (0.8) effect sizes were 9.56% [IQR 7.96%-11.5%], 34.45% [IQR24.61%-47.01%] and 70.41% [IQR 52.92%-85.91%]. None of these numbers reached the recommended minimal prospective power of 80%.

Figure 2 : True Report Probabilities



TRP curves derived from median prospective powers for 6 hypothetical effect sizes (Cohen's d : 0.1, 0.2, 0.5, 0.8, 1.2, 2). The orange curve indicates the TRP behaviour for a power of 80%. The green curve represents the TRP for the significant tests median prospective power. The green area represents the IQR. (number of tests = 90). A 50% hypothetical plausibility (prior probability) yielded TRPs of 44.2%, 72.4%, and 85.5% for small,medium and large effect sizes respectively. For a plausibility of 10% we found TRPs of 13.7%, 34.4%, and 54%.

Figure 3 : Overestimated effect sizes



Observed effect size for each positive test as a function of prospective power for 6 postulated effect sizes (partial êta squared: 0.00005, 0.019, 0.1, 0.195, 0.295, 0.4). Green dots represent effect sizes for t tests and orange dots for F tests. The dashed line shows the hypothetical effect size for power computation. Significant results yielded by underpowered tests tend to correspond to observed effect sizes greater than the (hypothetical) real effect size (Ioannidis, 2008). For a (hypothetical) small real effect size (0.019 or 0.1), more tests are underpowered and more significant tests overestimate the real effect size.

Discussion

These results generalize to a subfield of animal-model experimental psychopharmacoloy (nicotine CPP in mice) the lack of power frequently reported in the litterature of several neurobehavioural and psychological disciplines (Button et al., 2013; Maxwell, 2004; Schmidt-Pogoda et al., 2019; Walum et al., 2016). This can only accentuate the urgent need to enhance the methodological quality in that subflied, and likely also in the entire field, of animal-model experimental psychopharmacology.

Contact

francois.leonard@uliege.be

References

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated: *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Schmidt-Pogoda, A., Bonberg, N., Koecke, M. H. M., Strecker, J.-K., Wellmann, J., Bruckmann, N.-M., Beuker, C., Schäbitz, W.-R., Meuth, S. G., Wiendl, H., Minnerup, H., & Minnerup, J. (2019). Why most acute stroke studies are positive in animals but not in patients. *Annals of Neurology*, *n/a*(n/a). https://doi.org/10.1002/ana.25643

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00390

Walum, H., Waldman, I. D., & Young, L. J. (2016). Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies. *Biological Psychiatry*, *79*(3), 251–257. https://doi.org/10.1016/j.biopsych.2015.06.016