Neuroscience and Biobehavioral Reviews 84 (2018) 1–11

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Revisiting the validity of the mouse forced swim test: Systematic review and meta-analysis of the effects of prototypic antidepressants

N.Z. Kara[a,b], Y. Stukalin[a], H. Einat[a,b,c,*]

[a] School of Behavioral Sciences, Tel Aviv-Yaffo Academic College, Israel
[b] Department of Clinical Biochemistry and Pharmacology, Ben-Gurion University of the Negev, Israel
[c] College of Pharmacy, University of Minnesota, United States

ARTICLE INFO

Keywords:
Depression
Antidepressants
Animal models
Validity
Effect size
Q statistics

ABSTRACT

One problem area regarding animal models for affective disorders is unclear reproducibility, including external validity or generalizability. One way to evaluate external validity is with systematic reviews and meta-analyses. The current study presents a meta-analysis of the effects of prototypic antidepressants in the mouse forced swim test (FST).

We identified studies that examined effects of antidepressants in the FST in mice and used standard protocol, male mice and acute drug administration. We calculated Effect sizes using Cohen's d, homogeneity using Q statistic and correlations using Pearson's correlation.

Results indicate that all drugs reduce immobility in the FST. However, effect sizes for most drugs are heterogeneous and do not show a consistent dose/response relationship across variability factors. Reducing variability by examining only one strain or data from individual laboratories partially increases dose response relationship.

These findings suggest that whereas the FST is a valid tool to qualitatively screen antidepressant effects its validity in the context of hierarchical comparison between doses or compounds might be relevant only to single experiments.

1. Introduction

Animal models are essential for the study of human disease and for the development of better treatments. Animal models are frequently utilized in the study of neuropsychiatric diseases including affective disorders (Cryan and Slattery, 2007; Einat, 2007). Yet, the use of models in the research of affective disorders and their treatments is frequently debated with significant criticism. Some of the critics suggest that the models are not helpful enough in deciphering the underlying mechanisms of complex brain disorders and are not predictive enough to accurately anticipate drug effects in patients (Agid et al., 2007; Gould and Einat, 2007; Nestler and Hyman, 2010). One of the problem areas of models is unclear reproducibility, including both internal and external validity (Kafkafi et al., 2016). Low reproducibility rates are not unique to neuropsychiatric diseases models and a recent study suggests that within life science research the cumulative (total) prevalence of irreproducible preclinical research exceeds 50% of published results (Freedman et al., 2015). Reproducibility is also a major concern in human neuroscience and psychology research amounting to what had been termed in the last few years as the "replication crisis" (Aarts et al., 2015).

As recently indicated by Kafkafi and colleagues (Kafkafi et al., 2016), concerns regarding replicability and reproducibility in the field of mouse models related to psychobiology had been raised even before it became a broader concern in all areas of basic science. In mice behavioral work, the reproducibility is connected to several significant methodological challenges including the complex interaction between genotype and environment, defining and measuring proper behavioral constructs, and generalizability of mice models to human disease and disorders (Kafkafi et al., 2016).

One aspect of reproducibility is the external validity or generalizability, the extent to which results can be generalized when some factors in the experiments are changed. External validity can be evaluated by systematic replications where aspects of the experiment are manipulated in a controlled manner or by comparing effects in tests that are hypothesized to model similar states or traits such as in the execution of test batteries that are relevant to a specific disease (van der Staay et al., 2009). Alternatively, external validity can also be evaluated using systematic reviews and meta-analyses of available data, a standard practice in clinical research that is neglected in animal models

* Corresponding author at: School of Behavioral Sciences, Tel Aviv Yaffo Academic College, 14 Rabenu Yeruham St., Tel Aviv, Israel.
E-mail address: haimh@mta.ac.il (H. Einat).

https://doi.org/10.1016/j.neubiorev.2017.11.003
Received 26 April 2017; Received in revised form 1 November 2017; Accepted 1 November 2017
Available online 09 November 2017
0149-7634/

research. Compared with their frequent use in clinical research we found a very small number of such studies related to animal models of psychopathology. However, these few papers are very helpful in clarifying some important questions. For example, work by Jonasson (Jonasson, 2005) reviewed studies on sex differences in animal models of learning and memory and was able to identify areas of clear differences. Another recent meta-analysis of anxiety-related changes induced by sleep deprivation had very interesting results suggesting that standard anxiety tests for rodents are not translational to humans in the evaluation of sleep deprivation effects and that new tests and measures should be developed (Pires et al., 2016).

Clearly, many preclinical experiments are small, with less than ideal power, they are not systematically replicated and at time lack methodological rigor. These issues can lead to erroneous conclusions and very frequently to inflated effects. Recently we examined the power of some of the data we published from our own work and found high variability in the performance of mice across tests as well as a large variability in the power of different experiments. For example, we looked at experiments where we tested the effects of different compounds on the behavior of ICR male mice in the forced swim test (FST). The range of immobility time (the main measure of the test) for control (no interventions or treatments) mice across experiments was between mean of 122 s (Kara et al., 2016) and mean of 191 s (Kara et al., 2014). The range of power of these experiments was between 60% (Kara et al., 2016) and 85% (Sade et al., 2014). In an additional study that include interventions alongside lithium treatment, immobility time was at 195 s for the control group but power for the experiment was only between 20 and 30% (Toker et al., 2013). Such large differences, even in experiments that were performed by one research group, cast doubt on the external validity of the FST and raise a question regarding the possibility to compare results over different experiments. We suggest that conducting systematic reviews and meta-analyses of previously performed experiments can assist in clarifying the value and the limitations of animal models research. Such studies can help in selecting the most appropriate models for future research, increase the translational value of models and help in implementing the 3 R's of animal research ethics: replacement, reduction and refinement (de Vries et al., 2014; Sena et al., 2014). The importance of performing systematic reviews and meta-analyses had been recently emphasized (Hooijmans and Ritskes-Hoitinga, 2013).

In line with this idea, the present study presents a meta-analysis of the effects of prototypic antidepressants in the mouse forced swim test; one of the most frequently used screening models for antidepressant action. The study was designed to examine (1) can the FST qualitatively detect the effects of prototypic antidepressants across studies, conditions, mice strains, laboratories and time? (2) Can the FST be helpful in identifying hierarchical relationship of effects such as dose response relationship or strength of effects of different compounds across studies, conditional, strains, laboratories and time?

2. Methods

2.1. Search and selection of papers

First, we selected one representative drug from each antidepressant class. We selected imipramine, fluoxetine, bupropion and tranylcypromine from the tricyclic antidepressants, selective serotonin reuptake inhibitors, norepinephrine-dopamine reuptake inhibitors and monoamine oxidase inhibitors, respectively. Additionally, we searched for studies using the prototypic mood stabilizer lithium. These drugs were selected because they were documented to reduce mice immobility time in the FST and their common use in the study of antidepressant-like effects.

An exhaustive PubMed literature search was performed to identify studies that examined different antidepressant treatments on forced swim test behavior in mice. We used the key words forced swim test; Porsolt; mice; antidepressant; immobility; floating; behavioral despair; depression and animal model. Additionally; we searched the name of the drug or class of antidepressants specifically for each drug. The key search terms were deliberately kept broad in order to capture all potentially relevant papers.

For a study to be included in this meta-analysis, it had to meet a number of inclusion criteria. First, only experiments using a standard protocol in the FST were included. An experiment was included if (1) the test was conducted in the light phase; (2) trial duration was six minutes and (3) immobility time was scored in the last four minutes of each session. Other variations in the test were allowed. Second, only experiments using adult male mice were included and third, only experiments using acute intraperitoneal (IP) drug administration were included. One of the unique features of the FST is that it responds to acute treatment with antidepressant drugs and that these effects are augmented after chronic treatment (Cryan et al., 2005). Hence, the inclusion of acute and chronic experiments in one meta-analysis may not be appropriate. These inclusion criteria were designed to ensure consistency across experiments. We did not separate between scoring that was performed manually (as was in most of the older studies) or using automated systems (as in most of the newer studies) as there is enough data to show that both methods result in consistent results and that variability between automated and manual scoring is not larger than the variability between two scorers [e.g. (Crowley et al., 2004; Kurtuncu et al., 2005)].

Studies were selected in two phases. The first phase included screening of abstracts and titles according to the inclusion criteria. If necessary, full text manuscripts were consulted. The initial search identified twenty studies for imipramine, sixteen studies for fluoxetine, seven studies for bupropion, three studies for tranylcypromine and four studies for lithium. Studies were excluded mainly because their scoring included sessions of different lengths or that they used the rats' protocol (two exposures) in mice or because necessary data to calculate effect size were missing. The second phase included data extraction from full manuscripts.

Whenever a study did not report data in enough detail we tried to contact the authors and ask for the missing data. If contact attempts were unsuccessful, data was extracted directly from the graph or figure using a digital ruler. Studies comparing one control group to different dosage groups were analyzed as two or more studies. The reason for these separate analyses was the expectation that drug doses can influence treatment effects. Thus, in total, the final data came from 102 distinct experiments reported in 50 articles. In addition, to avoid heterogeneity derived from strain differences, studies were reanalyzed by mice strain and effect size was calculated separately for the most common strain in each drug. Last, we examined correlations between dose and effect for the entire cohort of studies and for sub-groups as detailed later.

2.2. Statistical analysis

Analyses were conducted separately for each drug and then again, within each drug for experiments that were conducted with the most common strain. Additional partial analyses were conducted for data coming from specific laboratories and experiments that were reported in one study.

Effect sizes were estimated using Cohen's d, an unbiased measure of the difference between two means (Cohen, 1992). Cohen's d is calculated by dividing the difference between the vehicle and active treatment groups by their pooled standard deviations.

Heterogeneity of effect sizes within each comparison was tested using the Cochran's Q test statistics (Higgins et al., 2002). A test for heterogeneity examines the null hypothesis that all studies are evaluating the same effect. The usual test statistic (Cochran's Q) is computed by summing the squared deviations of each study's estimate from the overall meta-analytic estimate, weighting each study's contribution in