Pharmacology, Biochemistry and Behavior xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Pharmacology, Biochemistry and Behavior

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

Analyzing test batteries in animal models of psychopathology with multivariate analysis of variance (MANOVA): One possible approach to increase external validity

Yelena Stukalin, Haim Einat*

School of Behavioral Sciences, Tel Aviv-Yaffo Academic College, Tel-Aviv, Israel

ARTICLE INFO

Keywords:
Test batteries
Affective disorders
Depression
Bipolar disorder
Animal models
Statistical analysis
Reproducibility

ABSTRACT

Background: One concern regarding animal models of psychopathology is unclear external validity. One way to establish external validity is to examine measures representing separate facets of the pathology with a battery of tests in the same cohort of animals. Additionally, utilizing the same animals in a battery of tests can help to reduce the number of animals in research. However, issues had been raised regarding the analysis of data coming from batteries and the standard practice is to analyze each test separately. This approach introduces two problems: (1) the analysis answers the question regarding separate tests but not regarding the general effect; (2) there is no correction for multiple comparisons. One way to overcome these challenges is to use transformations to *Z*-scores. We suggest an additional approach, analyzing test batteries with multivariate analysis of variance (MANOVA).

Methods: To compare the outcomes of *Z*-score analysis and MANOVA we re-analyzed two published studies where data were initially analyzed separately for each test. Additionally, we computed effect sizes.

Results: The first study tested interaction between sex and lithium in a battery of manic-like behaviors, the second study tested asenapine in a battery of anxiety-like behaviors. For the first study, the MANOVA analysis indicated no effects of sex and a significant antimanic-like effect of lithium and for the second study, the MANOVA indicated a significant anxiolytic effect of asenapine. *Z*-score analysis resulted in a significant general antimanic-like effect in the lithium study but failed to demonstrate the anxiolytic effects of asenapine in the second study.

Conclusions: It is possible to suggest that MANOVA is an appropriate way to analyze data from test batteries and that its use, when appropriate, can increase the validity, predictability and reproducibility of results.

1. Introduction

One important approach to study mechanisms of disease in general and neuropsychiatric illnesses in particular and to develop better treatments is through the utilization of appropriate animal models (McKinney, 2001; Valvassori et al., 2013). However, animal models are repeatedly criticized as not being helpful enough in deciphering the underlying mechanisms of these complex diseases and not predictive enough to accurately anticipate drug effects in patients (Nestler and Hyman, 2010; Gould and Einat, 2007). One concern is the unclear reproducibility of these models including both internal and external validity (Kafkafi et al., 2016; Fonio et al., 2012; Dzirasa and Covington, 2012; van der Staay et al., 2009).

One possible way to increase external validity is to look at a variety of behavioral measures that represent separate symptoms or separate domains of a disorder, symptoms and domains that may characterize different components of the pathology. This can be done while using distinct groups of animals for each behavioral test representing a symptom or a domain or it can be done with a battery of behavioral tests in the same cohort of animals. The older practice used to be with distinct groups. For example, a 1970's study was designed to explore the effects of lithium as a mood stabilizer and used separate groups of mice to examine lithium-induced reduction in the amphetamine-induced hyperactivity test and lithium's amelioration of reserpine-induced hypoactivity (Borison et al., 1978). The use of separate groups of mice for each test was standard practice but sometimes during the late 1980s this practice was becoming a highly limiting factor in many animal studies. Specifically, with the increasing ability to develop animals with targeted mutations and the utilization of such animals to decipher the biology of neuropsychiatric disorders and to identify drug targets, it

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

https://doi.org/10.1016/j.pbb.2017.11.003
Received 4 July 2017; Received in revised form 22 October 2017; Accepted 27 November 2017

became an issue to obtain enough animals for many separate groups. Most mice with targeted mutations were developed in academic laboratories with limited space and limited funds and many of these mutant animals were not very easy to breed. Hence, the need to use such mice for more than one test became critical. With that need, some researchers developed behavioral test batteries where the same animals were evaluated in a series of tests administered one after the other, sometimes continuously and in other cases with some time, hours or days separating between one test and the other. Significant issues were raised at the time regarding the possibility to use such behavioral batteries, their validity, and their specific structures. Fortunately, many of the conceptual and practical issues were answered with the seminal work of Paylor, Crawley and their colleagues who offered both a framework and practical solutions to many of the questions regarding these batteries (Crawley and Paylor, 1997; Crawley, 1999; Crawley, 2000; Bailey et al., 2006).

Further work revealed that beyond the necessity, there are also advantages to the utilization of behavioral test batteries compared with the use of separate groups for separate tests. Evaluation of the outcomes of an intervention be it environmental, pharmacological or molecular in a battery of tests comprising various tasks can provide strong support for a true positive disease-related or treatment-related phenotype. Moreover, a test battery can distinguish the effects of interventions across different behavioral tasks and therefore be helpful in making important connections between specific tasks and specific mechanisms (Cryan and Holmes, 2005). For example, within the arsenal of tests for anxiety-like behavior in rodents there are ways to evaluate state and trait anxiety, tests that are related to exploration, tests related to learned response, and tests related to conflict behavior (Cryan and Holmes, 2005; van Gaalen and Steckler, 2000); these tests can be used together to screen for an anxiolytic effect of a drug across the different domains (Cryan and Holmes, 2005). Accordingly, a drug that is designed to reduce learned fear in humans should be effective in tests that model learned responses whereas a general anxiolytic drug should be effective in all the different components. A complex battery of behavioral tests is therefore needed in order to answer such questions (Cryan and Holmes, 2005). A specific example in the context of bipolar disorder and mood stabilizing action is demonstrated in the work of Flaisher-Grinberg and colleagues with black Swiss mice (Flaisher-Grinberg and Einat, 2010). In a number of studies, these researchers suggested that the black Swiss strain of mice is a good model for behavioral domains of mania and in that context they examined whether the dissimilar mood stabilizers lithium and valproate can reduce manic-like behavior in a relevant battery of tests in black Swiss mice. The results of these studies showed that both drugs induce antimanic like changes but that they act on different behavioral facets of the model. The data indicated that both drugs reduced reward seeking behavior and amphetamine-induced hyperactivity but only valproate also reduced the heightened vigor of the mice in the forced swim test (Flaisher-Grinberg and Einat, 2010). Additional studies using a similar battery in this strain further demonstrated that the atypical antipsychotic risperidone reduced amphetamine-induced hyperactivity but did not influence behavior in the other tests of the battery (Hannah-Poquette et al., 2011) whereas the atypical antipsychotic asenapine reduced spontaneous activity, amphetamine-induced hyperactivity and vigor in the FST (Ene et al., 2015a). Hence, these studies, when taken together can suggest that the atypical antipsychotics are effective in the activity-related facets of the model whereas lithium and valproate induce effects that are beyond changes in activity.

Interestingly, the utilization of test batteries is a core component of clinical studies where in most cases disease state or the effects of interventions are evaluated across a number of domains, tasks and symptoms. For example, one of the standard ways to screen for antimanic drugs is by using the Young Mania Rating Scale (YMRS), a clinician administered questionnaire covering 11 different facets of mania including mood state, activity levels, sexual interest, sleep, irritability, speech, language/thought content, disruptive-aggressive behavior, appearance and insight (Young et al., 1978).

Last but not least, the utilization of the same animals in a battery of tests corresponds with the increasing attention to ethical principles and the expectation to reduce the number of animals used in research (Rowan, 1980).

Yet, the advantages of utilizing test batteries depend on appropriate analysis and interpretation of the data and some issues had been raised regarding the standard ways used to statistically analyze test batteries data. The objective of the current manuscript is to explore the possibility of utilizing multivariate analysis of variance (MANOVA) to analyze such batteries.

The standard practice in analyzing test batteries is to separately analyze each test in the battery using the appropriate statistics be it a Student's *t*-test if there are two groups, a one way ANOVA if there are more than two groups or a multivariate ANOVA if there are a number of factors interacting in each of the tests. For example, a recent study tested sex effects in Wistar Kyoto (WKY) rats, a strain that is known for susceptibility to depression-like behavior. This study used a battery of tests that included home-cage activity monitoring, open field test, marble burying test, novelty-induced hypophagia test, forced swim test (FST) and sucrose preferences test. The study also included a comparison of the WKY strain to an additional strain, Sprague Dawley (SD) rats. Animals were serially tested and data for each test were separately analyzed with two way ANOVA (with sex and strain as main factors) or with Kruskal-Wallis non-parametric ANOVA when data for the specific test were not homogeneous (Burke et al., 2016). The results of this study show that when analyzing each of the tests separately, there are differences between the sexes but only in some tests and not the others. Similarly, there are some differences between strains and some interactions between sex and strain (Burke et al., 2016). Another example for the standard practice of separate analysis of tests is demonstrated in the study of Lien and colleagues (Lien et al., 2008) who examined the behavioral consequences of a targeted mutation in the Bcl-2 gene in the context of affect. In this study, the researchers compared the behavior of Bcl-2 knockout mice with the wild type mice using a battery of tests that included spontaneous activity, sweet solution preference, elevated plus maze, resident-intruder aggression, forced swim test and amphetamine-induced hyperactivity. The data were analyzed separately for each test and the mutation was reported to significantly increase sweet solution preference and amphetamine-induced hyperactivity but had no effects in the other measures (Lien et al., 2008). There are two major problems with the results in both of these studies that may represent the conceptual issue of analyzing test batteries. First, the separate analysis of each test cannot answer the question regarding the general effect. The initial question of the first study was whether there are sex differences in the anxiety- and depression-like behaviors of WKY rats but the answer(s) are now test specific with no statistical conclusion for this main question. Similarly, in the second example, there is no general answer regarding the effects of the Bcl-2 mutation to induce manic-like behavior. (2) The battery in both studies included a number of tests (with some having more than one measure) and these tests were conducted in the same animals serially. Yet, because each test was analyzed in isolation, there was no attempt to correct for multiple comparisons and it is possible that some of the conclusions are exaggerated.

Unfortunately, the methodology of individual analysis of each separate test is the common approach and it is possible that in many cases it can result in either mistaken conclusions or inability to make general conclusions. We propose that the analysis of test batteries should be done in a way that will address these issues and therefore increase the external validity and predictability of these batteries. We further suggest that one statistical approach that may be utilized for the analysis of test batteries is multivariate analysis of variance (MANOVA). The MANOVA can be a helpful tool because it includes in one matrix all the tests of the battery. If the comprehensive MANOVA shows significant effects it supports a difference between groups across all test modalities.

The MANOVA can then be followed by appropriate post-hoc analysis for each of the separate tests in order to gain detailed information about the differences between groups.

One way that was previously suggested for the analysis of test batteries in animal models is with transformation of data to *Z*-scores followed by different methods of integrating *Z*-scores across tests (Huynh et al., 2011; Guilloux et al., 2011; Janus et al., 2015). For example, Guilloux et al. (2011) suggested averaging *Z*-scores across dependent variables and conducting a *t*-test for the independent variable. Similarly, Huynh and colleagues used a combined *Z*-score of two separate tests for depression-like behavior. Although the separate analysis of each individual test did not show a significant difference between groups, effects became significant after the combined *Z*-scores manipulation (Huynh et al., 2011). Whereas this method is not unreasonable, the MANOVA approach can give a stronger indication of the effects of an independent variable on the behavioral outcome in the entire battery. Moreover, using *Z*-scores does not eliminate the need for correction for multiple comparisons with the problematics that are attached to this type of tests. Additionally, expected effects in different tests may be in opposite direction (for example, an antidepressant is expected to reduce immobility in the forced swim test but increase social interaction) hence, simple averaging of *Z*-scores may reduce the possibility of identifying differences and additional data transformations are needed. However, using individual *Z*-scores may have an advantage for graphic presentation of results regardless of the method used for statistical analysis. Last, although MANOVA may be an appropriate way to analyze test batteries, it is important to remember that MANOVA relies on the assumption of linear relationship between dependent variables. However, with the type of tests used in most behavioral batteries such a relationship is assumed as a standard practice.

The present paper demonstrates the possibility of analyzing test batteries with MANOVA compared with analyzing with transformation to *Z*-Scores and of graphical presentation using individual *Z*-scores.

2. Methods

2.1. Using multivariate analysis of variance (MANOVA)

MANOVA is used in two major situations. The first is when there are several correlated dependent variables, and the researcher desires a single, overall statistical test on this set of variables instead of performing multiple individual tests. The second is to explore how independent variables influence some patterning of response on the dependent variables (Carey, 1998). The MANOVA gives one overall test of the equality of mean vectors for several groups. But it cannot tell which groups differ from which other groups on their mean vectors (Carey, 1998).

2.2. Averaged Z-scores

The *Z*-score is the number of standard deviations a data point is away from the mean. The *Z*-score is computed with a simple formula: $Z = (x - \mu) / \sigma$ where x is the individual value, μ is the population mean and σ is the standard deviation. For the current study we calculated the *Z*-score of the raw data separately for each test in the battery, averaged the *Z*-scores across tests and analyzed the resulting means using a one way ANOVA or a 2-way ANOVA as appropriate based on the independent factors.

2.3. Effect size

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 two groups by their pooled standard deviations.

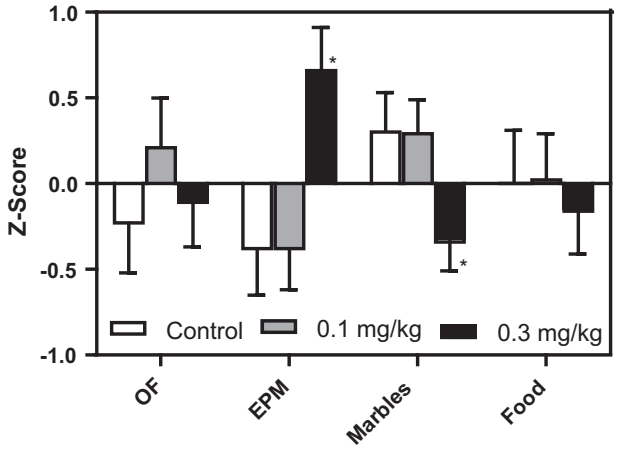

Fig. 1. Effects of asenapine in a battery of tests for anxiety-like behavior. Data presented as mean ± SEM of *Z*-scores. * represents statistically significant difference from control based on MANOVA and post-hoc analysis of the raw data. OF = Spontaneous activity in open field; EPM = open/closed arms time ratio in the elevated plus-maze; Marbels = number of marbles covered in the defensive marble burying test; Food = latency to eat in the novelty-induced hypophagia test.

3. Results

In both studies, the utilization of MANOVA to analyze the entire battery of tests in one matrix demonstrated a general effect of the intervention. Post-hoc analysis was efficient to further clarify the specific components of the battery where the effects of the intervention were significantly expressed. Analysis with averaged *Z*-scores resulted in a significant general effect in example 1 but failed to demonstrate the effects of intervention in example 2. This failure is the result of the opposite direction of the effects of the drug in the separate tests as can be seen with the graphical presentation (Fig. 1). Whereas we expect an anxiolytic drug to increase the open/closed time ratio in the EPM, we also expect it to decrease the number of marbles buried in the defensive marble burying test. The opposite direction of the effects resulted in the means of the *Z*-scores not to be different between the groups.

3.1. Example 1 – reanalysis of the effects of lithium and sex in a battery of tests in the black Swiss mouse model for mania-like behavior

We previously examined the effects of sex in the black Swiss mouse model, with and without lithium treatment and across a battery of behavioral tests (Ene et al., 2015b). The data were analyzed separately for each of the tests using 2-way ANOVA for most tests (with sex and lithium as main factors) and 3-way ANOVA for the amphetamine-induced hyperactivity test (with sex, lithium and amphetamine as main factors). The results of that analysis showed a significant effect of sex in one test (sweet solution preference) only and a significant effect of lithium in 4 out of 5 tests, spontaneous activity, sweet solution preference, EPM, and amphetamine-induced hyperactivity. We re-analyzed the data using both MANOVA and mean of *Z*-scores. The results of the MANOVA clearly indicate that for the entire study there was a significant effect of lithium [$F(5,31) = 7.53$, $p < 0.001$], no significant effect of sex [$F(5,31) = 0.82$, $p = 0.54$], and no interaction [$F(5,31) = 1.42$, $p = 0.25$]. Post hoc Bonferroni analysis shows significant effect of lithium in the spontaneous activity test ($p < 0.001$), the sweet solution preference test ($p < 0.002$), the EPM ($p < 0.001$) and the amphetamine-induced hyperactivity test ($p < 0.04$). Effect size estimation indicates strong effects in these measures with Cohen's d being 1.23 for spontaneous activity, 1.09 for sweet solution preference and 0.92 for the EPM. Effect size for the amphetamine-induced hyperactivity test was not computed because the design of this test does not allow such simple estimation. The analysis of the *Z*-score mean also

shows no effect of sex [$F(1,35) = 1.86$, $p = 0.18$], a significant effect of lithium [$F(1,35) = 8.72$, $p = 0.0006$] and no interaction [$F(1,35) = 1.18$, $p = 0.29$].

3.2. Example 2 – reanalysis of the anxiolytic-like effects of asenapine

We previously tested the effects of the atypical antipsychotic asenapine in a battery of tests for anxiety-like behavior. This study was conducted in two replications (Ene et al., 2015c). Separate, ANOVAs performed for each of the tests with asenapine and replication as main factors showed a significant anxiolytic-like effect of asenapine in the elevated plus-maze and the defensive marble burying test but not in the open field center time or the novelty induced hyponeophagia test (Ene et al., 2015c). We re-analyzed the data using MANOVA as well as using means of the *Z*-score. The results of the MANOVA clearly indicate an overall effect of asenapine [$F(8,66) = 2.48$, $p = 0.02$], an overall effect of replication [$F(4,33) = 4.59$, $p = 0.005$] and no interaction [$F(8,66) = 0.59$, $p = 0.78$]. Post hoc analyses (Bonferroni) show no significant effects of asenapine in the open field and the hyponeophagia tests but significant anxiolytic effects in the EPM [0.3 mg/kg ≠ saline control ($p < 0.03$)] and a trend for significant effect in the defensive marble burying test [0.3 mg/kg ≠ saline control ($p = 0.08$)]. Effect size estimation showed strong effects for these measures with Cohen's d being 1.0 for the EPM and 0.73 for the marble burying test. Interestingly, the analysis using means of *Z*-score did not result in a significant effect [$F(1,36) = 0.89$, $p = 0.35$].

4. Discussion

Preclinical studies, including ones utilizing animal models of psychopathology had been criticized for lack of reproducibility that is at least in part related to issues of external validity (Kafkafi et al., 2016; Fonio et al., 2012; Dzirasa and Covington, 2012; Aarts et al., 2015). One way to enhance external validity is to evaluate behavior across a number of domains that are related to the modeled phenomenon or to the symptoms or the domains of the modeled disorder. Indeed batteries of tests became a common approach in animal studies and are frequently used to examine interventions that result in pathology-like states (e.g. Varadarajulu et al., 2011; Kara and Einat, 2013) or to evaluate novel treatments (e.g. Kara et al., 2013; Kara et al., 2015). However, the general approach in analyzing such batteries is to treat each behavioral test independently of the others and analyze each one separately. This approach may be problematic as it is limited in the ability to make general conclusions across tests and it ignores the need for correction for multiple comparisons. These two issues are important as they reduce the ability to extrapolate from the results to possible effects in humans and they reduce the validity of the experiments. The current study demonstrates that using multivariate analysis of variance (MANOVA) to handle all data across tests could be an appropriate option for the statistical analysis of test batteries. The MANOVA analysis results in a general answer regarding the effects of the manipulation as it presents the probability that the manipulation induced a behavioral change across tests. Hence, the results of the test battery can indicate a general direction of the effect of a manipulation or a novel drug. For example, in our second example regarding the effects of asenapine in a battery of tests for anxiety-like behaviors, the separate analysis of the different tests permitted statements such as “asenapine induced an anxiolytic-like behavioral change in tests A and B but not test C and D”. The MANOVA analysis now permits the general statement that asenapine induced an anxiolytic-like effect and only the post hoc analysis will clarify in which of the tests the effect was statistically significant or not. Additionally, the effects indicated by the MANOVA have a lower probability for false positive conclusions due to lack of corrections for multiple comparisons in the standard practice. It is true that the issue of multiple comparisons in test batteries can be addressed with the appropriate corrections but unfortunately, in most studies the corrections are not applied. Interestingly, whereas we did not find studies that used MANOVA to analyze test batteries in animal models of psychopathology, this statistical method was utilized in the analysis of test batteries in humans (Schatz et al., 2006). It is however important to note that after performing a MANOVA, it is essential to select the appropriate post-hoc comparison methods that account for multiple comparisons. Most of the tests that are included in the major statistical analysis software today offer such tools including the Bonferroni post-hoc test, Scheffe post-hoc test, Sidak test, Tukey test and others (McHugh, 2011; Liu et al., 2010).

Despite the advantages of the MANOVA, it is possible that when a battery of tests includes sub groups from different realms, the MANOVA approach may not be the most appropriate analysis. For example, in many studies of mice with targeted mutations there is a need to perform tests for basic neuromotor behaviors before performing specific tests in the context of the research question as the neuromotor function is a prerequisite to meaningful interpretation of other behavioral tests (Crawley, 1999). For example, Frederick and colleagues (Frederick et al., 2012) performed a neurobehavioral assessment of the G(αq) receptor knockout mouse and included motor coordination, motor activity, cognitive, anxiety and depression related testing. In such studies, the performance in some tests depends on basic motor abilities and it is therefore sensible to first analyze the neuromotor components separately before constructing a MANOVA matrix to analyze the behavioral tests related to other domains.

In summary, we suggest that multivariate analysis of variance (MANOVA) followed by suitable post-hoc tests, is an appropriate way to analyze significant portion of the data from behavioral test batteries in animal models of neuropsychiatric disorders. We suggest that using MANOVAs will increase the validity, the predictability and the reproducibility of results.

Acknowledgement

Study was partially supported by a grant in aid of research to HE from the Tel Aviv-Yaffo Academic College.

References

Aarts, A.A., et al., 2015. Estimating the reproducibility of psychological science. Science 349 acc4716.

Bailey, K.R., Rustay, N.R., Crawley, J.N., 2006. Behavioral phenotyping of transgenic and knockout mice: practical concerns and potential pitfalls. ILAR J. 47 (2), 124–131.

Borison, R.L., et al., 1978. Lithium prevention of amphetamine-induced ‘manic’ excitement and of reserpine-induced ‘depression’ in mice: possible role of 2-phenylethylamine. Psychopharmacology 59 (3), 259–262.

Burke, N.N., et al., 2016. Sex differences and similarities in depressive- and anxiety-like behaviour in the Wistar-Kyoto rat. Physiol. Behav. 167, 28–34.

Carey, G., 1998. Multivariate Analysis of Variance (MANOVA): I. Theory. Academic Press, Boston.

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159.

Crawley, J.N., 1999. Behavioral phenotyping of transgenic and knockout mice: experimental design and evaluation of general health, sensory functions, motor abilities, and specific behavioral tests. Brain Res. 835 (1), 18–26.

Crawley, J.N., 2000. What's Wrong With My Mouse?: Behavioral Phenotyping of Transgenic and Knockout Mice, 1 ed. Wiley-Liss, New York, pp. 376.

Crawley, J.N., Paylor, R., 1997. A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice. Horm. Behav. 31 (3), 197–211.

Cryan, J.F., Holmes, A., 2005. The ascent of mouse: advances in modelling human depression and anxiety. Nat. Rev. Drug Discov. 4 (9), 775–790.

Dzirasa, K., Covington 3rd, H.E., 2012. Increasing the validity of experimental models for depression. Ann. N. Y. Acad. Sci. 1265, 36–45. http://dx.doi.org/10.1111/j.1749-6632.2012.06669.x.

Ene, H.M., Kara, N.Z., Einat, H., 2015a. The effects of the atypical antipsychotic asenapine in a strain-specific battery of tests for mania-like behaviors. Behav. Pharmacol. 26, 331–337.

Ene, H.M., Kara, N.Z., Einat, H., 2015b. Introducing female black Swiss mice: minimal effects of sex in a strain-specific battery of tests for mania-like behavior and response to lithium. Pharmacology 95 (5–6), 224–228.

Ene, H.M., et al., 2015c. Effects of repeated asenapine in a battery of tests for anxiety-like behaviours in mice. Acta Neuropsychiatr. 11, 1–7.

Flaisher-Grinberg, S., Einat, H., 2010. Strain specific battery of tests for separate behavioral domains of mania. Front. Psych. 1 (10), 1–10.

Fonio, E., Golani, I., Benjamini, Y., 2012. Measuring behavior of animal models: faults and remedies. Nat. Methods 9 (12), 1167–1170.

Frederick, A.L., Saborido, T.P., Stanwood, G.D., 2012. Neurobehavioral phenotyping of G(alphaq) knockout mice reveals impairments in motor functions and spatial working memory without changes in anxiety or behavioral despair. Front. Behav. Neurosci. 6, 29.

van Gaalen, M.M., Steckler, T., 2000. Behavioural analysis of four mouse strains in an anxiety test battery. Behav. Brain Res. 115 (1), 95–106.

Gould, T.D., Einat, H., 2007. Animal models of bipolar disorder and mood stabilizer efficacy: a critical need for improvement. Neurosci. Biobehav. Rev. 31 (6), 825–831.

Guilloux, J.P., et al., 2011. Integrated behavioral z-scoring increases the sensitivity and reliability of behavioral phenotyping in mice: relevance to emotionality and sex. J. Neurosci. Methods 197 (1), 21–31.

Hannah-Poquette, C., et al., 2011. Modeling mania: further validation for Black Swiss mice as model animals. Behav. Brain Res. 223 (1), 222–226.

Huynh, T.N., et al., 2011. Sex differences and phase of light cycle modify chronic stress effects on anxiety and depressive-like behavior. Behav. Brain Res. 222 (1), 212–222.

Janus, C., et al., 2015. Behavioral abnormalities in APPSwe/PS1dE9 mouse model of AD-like pathology: comparative analysis across multiple behavioral domains. Neurobiol. Aging 36 (9), 2519–2532.

Kafkafi, N., et al., 2016. Reproducibility and replicability of rodent phenotyping in preclinical studies. In: bioRxiv.

Kara, N.Z., Einat, H., 2013. Rodent models for mania: practical approaches. Cell Tissue Res. 354 (1), 191–201.

Kara, N.Z., et al., 2013. Trehalose induced antidepressant-like effects and autophagy enhancement in mice. Psychopharmacology 229 (2), 367–375.

Kara, N.Z., Flaisher-Grinberg, S., Einat, H., 2015. Partial effects of the AMPAkine CX717 in a strain specific battery of tests for manic-like behavior in black Swiss mice. Pharmacol. Rep. 67 (5), 928–933.

Lien, R., et al., 2008. Behavioral effects of Bcl-2 deficiency: implications for affective disorders. Pharmacol. Rep. 60 (4), 490–498.

Liu, C., Cripe, T.P., Kim, M.O., 2010. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. Mol. Ther. 18 (9), 1724–1730.

McHugh, M.L., 2011. Multiple comparison analysis testing in ANOVA. Biochem. Med. (Zagreb) 21 (3), 203–209.

McKinney, W.T., 2001. Overview of the past contributions of animal models and their changing place in psychiatry. Semin. Clin. Neuropsychiatry 6 (1), 68–78.

Nestler, E.J., Hyman, S.E., 2010. Animal models of neuropsychiatric disorders. Nat. Neurosci. 13 (10), 1161–1169.

Rowan, A.N., 1980. The concept of the three R's. An introduction. Dev. Biol. Stand. 45, 175–180.

Schatz, P., et al., 2006. Sensitivity and specificity of the ImPACT Test Battery for concussion in athletes. Arch. Clin. Neuropsychol. 21 (1), 91–99.

van der Staay, F.J., Arndt, S.S., Nordquist, R.E., 2009. Evaluation of animal models of neurobehavioral disorders. Behav. Brain Funct. 5, 11.

Valvassori, S.S., et al., 2013. Contributions of animal models to the study of mood disorders. Rev. Bras. Psiquiatr. 35 (Suppl. 2), S121–31.

Varadarajulu, J., et al., 2011. Increased anxiety-related behaviour in Hint1 knockout mice. Behav. Brain Res. 220 (2), 305–311.

Young, R.C., et al., 1978. A rating scale for mania: reliability, validity and sensitivity. Br. J. Psychiatry 133, 429–435.