Reliability of Mouse Behavioural Tests of Anxiety: a Systematic Review and Meta-Analysis on the Effects of Anxiolytics.

Marianna Rosso, Robin Wirz, Ariane Vera Loretan, Nicole Alessandra Sutter, Charlène Tatiana Pereira da Cunha, Ivana Jaric, Hanno Würbel, Bernhard Voelkl*

Division of Animal Welfare, University of Bern, Bern, Switzerland

**corresponding author: marianna.rosso@vetsuisse.unibe.ch*

ABSTRACT

Animal research on anxiety and anxiety disorders relies on valid animal models of anxiety. However, the validity of widely used rodent behavioural tests of anxiety has repeatedly been questioned, as they often fail to produce consistent results across independent replicate studies using different study populations or different anxiolytic compounds. In this study, we assessed the sensitivity of behavioural tests of anxiety in mice to detect anxiolytic effects of drugs prescribed to treat anxiety in humans. To this end, we conducted a pre-registered systematic review of studies reporting tests of anxiolytic compounds against a control treatment using common behavioural tests of anxiety in mice. PubMed and EMBASE were searched on August 21st 2019 for studies published in English and 814 papers were identified for inclusion. Risk of bias was assessed based on Syrcle's risk of bias tool and the Camarades study quality checklist on a randomly selected subsample of 180 papers. Meta-analyses on effect sizes of treatments using standardized mean differences (Hedges' g) showed that only two of 17 test measures reliably detected effects of anxiolytic compounds other than diazepam. Further, we report considerable variation in both direction and size of effects of most anxiolytics on most outcome variables, indicating poor replicability of test results. This was corroborated by high heterogeneity in most test measures. Finally, we found an overall high risk of bias. Our findings indicate a general lack of sensitivity of common behavioural tests of anxiety in mice to anxiolytic compounds and cast serious doubt on both construct and predictive validity of most of those tests. The use of animals to model human conditions can be justified only if the expected results are informative, reproducible, and translatable. In view of scientifically valid and ethically responsible research, we call for a revision of behavioural tests of anxiety in mice and the development of more predictive tests.

INTRODUCTION

Animal experiments are a key component of basic and preclinical research, where the mechanisms of diseases are studied and new compounds for their treatment are examined for safety and efficacy before being tested in humans (fda.gov). However, the use of animals for research can only be justified when the results obtained are informative (1–3), replicable* (4–6), and translatable* (7,8). Furthermore, public concern for animal welfare urges scientists to comply with the 3Rs principle (9), that is to refine, reduce, or replace the use of animals whenever possible (10,11).

To achieve these goals and ensure responsible scientific practice, the validity* of animal models in use is pivotal (2,12–14). A growing body of evidence indicates the lack of validity of animal models as a potential cause for translational failure (13,15–17). Translational failure can slow down medical advancement in the treatment of human disorders (18–20), put patients in clinical trials at risk (3), waste research resources (21), and harm animals for inconclusive research.

Anxiety disorders are amongst the most common mental health conditions, requiring still new and better treatments (22–26). To study anxiety and to test the efficacy of anxiolytic compounds behavioural tests in mice and other animals are commonly used (22,23,27,28). Such tests are mostly based on exploiting an approach-avoidance conflict, i.e. the conflict an animal may experience between exploring a new, and avoiding a potentially threatening, environment (27,29,30). Amongst the various behavioural tests for rodents, the open-field test is arguably the most popular one (23). This test, although with several modifications (31,32), generally consists of a brightly illuminated arena, enclosed by walls. During the test, an animal is placed inside the arena and behavioural outcomes are recorded. The test was originally established to assess emotionality in rats, using urination and defecation as measures of timidity (31,33). The use of the open-field test was then extended to assess a wider range of behavioural features and psychiatric conditions (27) and adopted for other species. Similar to rats, early studies which employed the open-field test in mice measured

defecation and freezing to assess genetic differences in behaviour (34,35). Additionally, the distance travelled in the open-field test has been introduced and--since then--widely used as a measure of locomotor activity to assess, for instance, the effect of sedative or stimulant drugs (36). Further, thigmotaxis in the open-field, namely the tendency to explore the proximity of the walls while avoiding the centre of the arena, is often recorded and interpreted as a proxy for anxiety (27,32,37).

Similar to the open-field test, the elevated plus maze test (38) and the light-dark box test (39) are based on the conflict between the exploration of a new environment and the natural aversion of rodents to bright and open spaces. The rationale behind these tests as measures of anxiety rests on the assumption that a state of anxiety should modulate the animals' behaviour by reducing exploration, therefore reducing the exposure to (potential) threats (22,27,40). Accordingly, the efficacy of anxiolytic compounds is assessed based on whether and to what extent they attenuate the reduction of exploratory behaviour by the test situation. Other popular tests, such as the hole-board test (41), the elevated zero maze (42), the social interaction test (43), the novelty suppressed feeding test (44), and the four-plate test (45), are based on the same conceptual rationale.

Over the years, behavioural tests for anxiety have been considered validated, because of reported behavioural changes elicited by benzodiazepines, and specifically diazepam (46–48). However, anxiolytic agents such as benzodiazepines also possess anti-depressant and sedative effects, which implies that the observed behavioural effects may not necessarily be due to a change in anxiety, but could be a result of the sedative properties of the drug (36).

Despite their popularity, several experimental studies, as well as literature reviews, have highlighted inconsistent results in the behavioural outcomes elicited by new classes of anxiolytics, therefore questioning the suitability of these outcomes as indicators for anxiety (29,36,46,49,50). Benzodiazepines, although popular in the past to treat anxiety, have now been replaced by better pharmacological compounds with fewer side effects and lower withdrawal-related risks (51–53). Selective Serotonin Reuptake Inhibitors (SSRIs) or

Serotonin–Norepinephrine Reuptake Inhibitors (SNRIs), which are now used as a first-line pharmacological treatment for human anxiety disorders, have failed to give reliable results in rodent behavioural tests of anxiety (29,36,46,50,54).

Here, we aimed to assess the validity of common behavioural tests of anxiety in mice by evaluating their responsiveness to anxiolytic compounds prescribed to humans, a process known as 'reverse translation' (55,56). To this end, we performed a pre-registered systematic review of research papers that had used these tests on laboratory mice, for a broad range of anxiolytic compounds. We investigated the overall effect size for a range of test measures of common behavioural tests as well as the variation of the reported outcomes across the published literature. Additionally, we evaluated sample heterogeneity and estimated the quality of reporting through a risk of bias assessment.

***Glossary of key terms**

1. **Replicability**: the likelihood with which results can be replicated by an independent study.
    - Relevant literature: (5,6,57–61)
2. **Translatability**: the extent to which results obtained in an animal model can be replicated in the system which is being modelled.
    - Relevant literature: (16–18,62–64)
3. **Validity**: to be fit for use in research, and therefore be considered to be a valid animal model, a test or animal model should meet several criteria of validity, including:
    i. **Construct validity**: the extent to which the test can measure what it is supposed to measure
    ii. **Predictive validity**: the extent to which a test can predict a certain outcome in the system that is being modelled.
    - Relevant literature: (1,2,12,28,65,66)

METHODS

PRE-REGISTRATION.

Prior to data extraction, in November 2019, this study was pre-registered at SYRCLE (see supplementary information for the pre-registration protocol).

SEARCH STRATEGY.

The search strategy consisted of i) a list of anxiolytic compounds, ii) the keyword “mice”, and iii) a list of behavioural tests for anxiety. To define the list of anxiolytic compounds, we used a combination of the following databases to list compounds that are commonly used to treat anxiety disorders in humans: DrugBank (drugbank.ca); FDA Drug Approval Databases (fda.gov); Anxiety and Depression American Association (adaa.org). We selected the following compounds: alprazolam, amitriptyline, buspirone, chlordiazepoxide, citalopram, clomipramine, clonazepam, clorazepate, desipramine, diazepam, doxepin, duloxetine, escitalopram, fluoxetine, flurazepam, fluvoxamine, hydroxyzine, imipramine, lorazepam, maprotiline, mirtazapine, nortriptyline, oxazepam, paroxetine, protriptyline, sertraline, temazepam, trazodone, triazolam, trimipramine, venlafaxine. A literature search allowed us to identify behavioural tests commonly used to assess anxiety in mice (Table 1). Each test that yielded more than 10 results, when searched on PubMed (on date July 15th 2019) in combination with the aforementioned list of compounds, and the keyword “mice”, was included in the search (Supplement 1). The search was performed on PubMed (ncbi.nlm.nih.gov/pubmed) and EMBASE (embase.com), on August 21st, 2019.

Test	Test measure	N Outcomes retrieved	Included
Elevated plus maze (EPM)	eca: Number of entries into closed arms.	206	yes
	eoa: Number of entries into open arms.	296	yes
	toa: Time (both in percentage and in time unit) spent in the open arms.	552	yes
Elevated zero maze (EZM)	ecc: Number of entries into the closed compartment.	2	no
	eoc: Number of entries into the open compartment.	5	no
	toc: Time (both in percentage and in time unit) spent in the open compartment.	14	yes
Four-plate test (FPT)	cross: Number of punished crossings.	42	yes
Holeboard test (HBT)	hd: Number of head dips.	137	yes
Light-dark box (LDB)	dark: Time spent in the dark compartment.	35	yes
	light: Time (both in percentage and in time unit) spent in the light compartment.	187	yes
	trans: Number of transitions between the two compartments.	107	yes
Novelty suppressed feeding (NSF)	lat: Latency to eat (sec).	37	yes
Open field test (OF)	cent: Time (both in percentage and in time unit) spent in the center (as defined by the authors).	87	yes
	dist: Distance travelled.	125	yes
	rear: Number of rearings.	207	yes
	sqrs: Number of squared crossed.	362	yes
Social interaction test (SI)	time: Time (sec) spent in social interaction.	26	yes
Staircase test (STC)	rrs: Number of rearings.	27	yes
	stps: Number of steps climed.	29	yes
Vogel conflict test (VC)	dbs: Number of drinking bouts.	7	no
	shck: Number of shocks accepted or received.	9	no

Table 1 – Behavioural tests for anxiety in mice and relative test measures included in the search.

STUDY SELECTION

After reference retrieval, we excluded paper duplicates using the reference manager software Citavi 6.4 (Swiss Academic Software GmbH, Wädenswil, CH). The main reviewer (MR) scanned the titles, abstracts and/or methods of these papers, and excluded all those, which did not use the behavioural tests of interest (Table 1), mice, or the selected anxiolytic

compounds. Additionally, we excluded papers that were not original research papers and papers that were not written in English. After the first scan, two independent reviewers (main reviewer: MR, second reviewers: RW, AL, NS) performed the full paper screening and the data extraction.

STUDY CHARACTERISTICS

Studies were included or excluded according to the pre-specified inclusion/exclusion criteria (Supplement 1). For each paper, two reviewers independently extracted information about the animals (i. strain, ii. sex, iii. age, iv. transgenic ID; v. stress or defeat treatment), about the treatment (vi. compound, vii. dosage, viii. route of administration, ix. time of administration before testing), and about testing (x. open field size, xi. test duration). For each test, we selected test measures suggested by the authors as measures of anxiety (Table 1). For each test measure, we extracted mean values, sample size, and either standard deviation or standard error of the mean, for both treatment and control group. We accepted any control group as declared by the authors (e.g. administrating water, saline solution, etc.). Information from graphical data was extracted using the online software Automeris (https://apps.automeris.io/wpd/).

DATA ANALYSIS

The statistical analysis was performed in R (1.4.1103) (67) with the package metafor 2.4-0 (68). For each study, we computed the standardized mean difference Hedges' g between the control and the treatment group as the chosen indication of effect size (metafor::escalc). We included any test measure that yielded at least 10 results. Consequently, four measures (EZM-*eoc*, EZM-*ecc*, VT-*shcks*, VT-*dbs*) were excluded from further analysis. For the measures LDB-*dark*, EPM-*eca*, NSF-*lat*, STC-*rrs* we reversed the sign of the effect size, because a decrease in behaviour manifestation is expected as a result of treatment. Our data pool was subset by test measure and a meta-regression model was fitted for every subset.

rma (yi, vi, mods= ~ factor (compound) - 1, random = list(~ 1 | study/observation,~ 1 | strain)

Standardized mean differences (Hedges' g) were tested with the modifier 'compound' (anxiolytic compounds) against the null hypothesis of the estimated effect size for each compound group equalling zero. Publication and strain were added as random effects. To assess the overall estimated effect size, independent of anxiolytic compound, the same model syntax was used, excluding the factor modifier. Total and partial I^2, indicating the percentage of sample variation, were used as a measure of heterogeneity, and were calculated using the methods proposed in (69).

RISK OF BIAS

Due to the large sample size, an assessment of quality was made on a subsample consisting of 180 randomly selected papers. The assessment was done by two independent reviewers (MR, CP), who evaluated 80 different papers each, as well as 20 papers that were reviewed by both investigators, to estimate inter-rater reliability. We used an adapted combination of the CAMARADES study quality checklist and SYRCLE's risk of bias tool (Supplement 1).

RESULTS

STUDY SELECTION

Our search retrieved 744 papers from PubMed and 2533 papers from EMBASE of which 1764 were excluded in the first steps of the review (Fig 1). In particular, 533 were excluded as paper duplicates, and 1231 were excluded based on abstract and/or method section screening. The full texts of 1513 papers were screened and 814 of those papers were included in the data extraction process according to the pre-specified criteria. As the search strategy identified key words in all fields of the text, several papers not relevant to us were identified; 331 papers were excluded because the sample size was unclear or not reported, 62 papers were excluded because the text was unavailable publicly, 59 papers were excluded because compounds other than the ones of interest were used, or compounds were used in combination with other compounds, 48 papers were excluded because of issues in the reporting of the outcomes, 40 papers were excluded because they had formats other than research papers, 33 papers were excluded because the behavioural tests used were different from the ones of interest, 25 papers were excluded due to ambiguity regarding the measure of variance of the reported outcomes, 24 papers were excluded because they used animals other than mice, or because of ambiguity in the species of animal used, and 13 papers were excluded for other reasons (i.e. missing controls, treatment administered to mothers, etc.).

STUDY CHARACTERISTICS

All the eligible studies used mice, which were tested in behavioural tests after administration of anxiolytic compounds. The Supplementary table illustrates the details of data distribution in the different test measures of interest in combination with each compound. Due to reporting of multiple outcomes per paper, a total of 2476 outcomes were distributed across 17 different test measures, in combination with 25 different anxiolytic compounds. The test measures from the elevated plus maze and the open field made up the great majority of outcomes (74%, Table 1), followed by the light-dark box test and the holeboard test

contributing a total of 13% and 5% of the outcomes, respectively. A minor contribution was attributed by the staircase test (the staircase test, n = 56, “rrs” n = 27, “stps” n = 29), the four-plate test (n = 42), the novelty suppressed feeding test (n = 37), the social interaction test (n = 26), and the elevated zero maze (n = 14). The great majority of these measures were recorded when used in combination with benzodiazepines (72%), with diazepam being the most frequently used compound (65%). SSRIs was the second most common compound class (20%), with fluoxetine (12%) being its most frequently used representative.

RISK OF BIAS

A sub-sample of 180 papers was analysed in detail to assess the risk of bias across 17 different items (Table 2). All the scored papers were published in peer-reviewed journals, and most of them reported mouse strain (95%), sex (90%) and housing temperature (75%). 31% of the papers reported details regarding compliance with animal welfare regulations, 43% of the papers reported details on the statistical analysis, and 34% of the papers reported details on the blinding procedures. For the following five items, we scored a high risk of bias: automatic allocation to treatment group (97%), randomized order of testing (92%), a-priori sample size calculation (98%), random housing (95%), and blinding of investigators (95%). Further details are reported in Table 2.

Question	High	Medium	Low
was an automatic randomization method used to allocate animals to groups?	97.22	2.78	0
were animals randomly allocated to treatment/control group?	65.56	34.44	0
was the test order randomized or counterbalanced?	92.78	6.11	1.11
was the sample size declared to be appropriately calculated?	98.89	1.11	0
where animals randomly housed?	95.56	4.44	0
compliance with animal welfare regulations declared?	19.44	48.89	31.67
were the investigators blinded during the experiment?	95.56	3.89	0.56
is the statistical analysis described?	2.22	54.44	43.33
Is the housing temperature reported?	25	0	75
Is the sex of the animals reported?	10	0	90

Is the strain of the animals reported?	5	0	95
conflict of interest declaration	52.78	0	47.22
publication in a peer-reviewed journal?	0	0	100
were the outcome assessors blinded during the experiment?	65.56	0	34.44

234

235 **Table 2**: Results of the risk of bias assessment. Values in the table indicate percentages of
236 papers, which scored either as high, medium, or low risk of bias in each item (row).

237

SYNTHESIS OF RESULTS

Estimated effect sizes varied greatly across the majority of the test measures and compounds (Fig 2). The overall estimated effect size allows determining whether there is evidence of an anxiolytic effect on the behavioural measures elicited by a range of anxiolytic compounds. Ten out of the 17 test measures yielded a positive overall effect size significantly different from zero (EPM-*eca*, EPM-*eoa*, EPM-*toa*, FPT-*cross*, LDB-*light*, LDB-*trans*, NSF-*lat*, OF-*cent*, SI-*time*, STC-*rrs*), while overall effects of the remaining seven did not significantly deviate from zero.

For each meta-analysis, the factor 'compound' was tested for significance to assess whether any of the anxiolytic compounds affected behavioural outcomes. For this, the null hypothesis to be tested assumes the estimated effect sizes for all compounds to be zero (68). After family-wise correction for multiple testing for the 17 meta-analyses performed, five measures showed no significant effect, namely EZM-*toc*, LDB-*dark*, NSF-*lat*, OF-*dist*, and SI-*time* (Table 3).

For each test measure, we calculated total and partial I^2 as a measure of heterogeneity. For 15 out of 17 measures, total I^2 was above 85%. The partial I^2 attributed to 'strain' contributed little to the total I^2, except for SI-*time* , where it accounted for 48% of the total heterogeneity. Partial I^2 attributable to within-study heterogeneity varied greatly across measures: in 10 cases being <10%, while being more pronounced in others (e.g. 64% for FPT-*cross*). Between-study heterogeneity explained the greater part of the total heterogeneity for 14 out of the 17 measures (Table 3).

Given the 25 compounds and 17 test measures, there are a total of 425 compound-by-measure combinations. We found reported study outcomes for 182 of those compound-by-measure combinations (details summarized in the Supplementary Table). The number of outcomes per combination varied from 1 to 413, with 118 compound-measure combinations with more than one outcome recorded. Of these, only 32 had a positive and significant effect size (i.e. the lower bound of the 95% confidence interval being larger than zero), while 86

combinations did not show a positive effect (Fig 2 and Supplementary Table). Diazepam was the compound that elicited a significant positive effect size in 9 out of 17 test measures. Overall, most of the combinations with a significant effect size were due to benzodiazepines, with 20 positive effects out of 32. LDB-*light* yielded a positive effect size for most of the anxiolytic compounds tested, 8 out of 11, and EPM-*toa* yielded a positive effect size for 5 out of 15 anxiolytic compounds. The rest of the test measures detected an effect for at most two anxiolytic compounds, across the range with which they were tested.

The percentage of individual observations that detected a positive significant effect varied greatly across the different combinations of test measures and anxiolytic compounds, ranging from 0% to 100% (Table 4). As all the compounds included in this analysis have been shown to reduce anxiety in humans, we assessed the sensitivity of behavioural tests outcomes to detect the expected anxiolytic effect of these compounds in mice based on the logic of reverse translation. Thus, we used the proportion of individual studies reporting a significant positive effect as a measure of sensitivity and an estimate of the true positive rate. To conclude that a behavioural test reliably detects an anxiolytic effect, we require that individual studies detect significant effects (positive effect size with a 95% confidence interval not including zero) in at least three out of four cases (i.e. 75%). The majority of behavioural measures failed to reliability detect an effect for the majority of the compounds. In 89 out of 118 combinations for which more than one outcome was recorded, less than 75% of individual studies reported significant positive effects, while only for 29 combinations, the proportion was greater than 75%. Table 4 suggests that diazepam was the compound that most often elicited a behavioural change detectable in five test measures. Here, we also observe a higher number of studies as compared to other compounds. Out of the 29 'reliable' combinations, benzodiazepines were the dominant compound class, showing reliable results in 14 combinations. LDB-*light* seems to be the most promising candidate to detect an anxiolytic effect, with the majority of individual studies detecting an effect in seven out of 11 anxiolytic compounds across compound classes. Furthermore, EPM-*eoa* and EPM-

toa reliably detected effects for 3 and 4 anxiolytic compounds, respectively. Similarly, OF-*sqrs*, reliably detected an effect of 3 anxiolytic compounds, but the number of individual studies was far lower than for the EPM. Forest plots (Fig 3 and Supplementary Material) show how for some measures the estimated effect sizes for individual studies range from highly negative values to highly positive ones, spreading in an almost symmetrical fashion across the null. Clear examples of such pattern can be seen in the forest plots of EPM-*eca*, HBT-*hd*, LDB-*trans*, NSF-*lat*, OF-*dist*, OF-*rear*, OF-*sqrs*, and STC-*stps*.

Test	Measure	Significance of factor 'compound'	I^2 Total	I^2 between studies	I^2 within study	I^2 Strain
EPM	eca	*	90.3	84.4	5.5	0.4
	eoa	*	87.4	57.3	9.1	21
	toa	*	94.3	73.5	4.5	16.4
EZM	toc	ns	85.3	0	0	85.3
FPT	cross	*	85.5	21.5	64	0
HBT	hd	*	97.7	97.7	0	0
LDB	dark	ns	99.2	99.1	0.1	0
	light	*	96.2	92.4	0.7	3.2
	trans	*	69.4	64.7	0	4.7
NSF	lat	ns	91.8	54.9	36.9	0
OF	cent	*	90.2	77.9	11.1	1.1
	dist	ns	82.9	57	19.4	6.5
	rear	*	93.4	91.3	2.1	0
	sqrs	*	95.1	85.9	8.4	0.8
SI	time	ns	94.6	0	45.9	48.7
STC	rrs	*	86.2	60	26.2	0
	stps	*	97.1	78.5	0	18.6

Table 3: Significance level of moderator effect (treatment × compounds interaction), total and partial I^2 estimates per test measure.

			Compounds													
			Benzodiazepine					Other	SNRI		SSRI					
Test	Measure		alprazolam	chlordiazepoxide	clorazepate	diazepam	lorazepam	hydroxyzine	duloxetine	venlafaxine	buspirone	citalopram	escitalopram	fluoxetine		
EPM	eca	*n*		21		138		2	2		22		2	13		
		%		5%		38%		100%	0%		32%		0%	8%		
		sign.														
	eoa	*n*	4	17		221		2	2		25			14		
		%	75%	59%		82%		100%	100%		20%			21%		
		sign.														
	toa	*n*	8	24		413	2	3	4	3	32	3	5	35		
		%	75%	71%		84%	50%	100%	0%	33%	34%	0%	40%	23%		
		sign.														
EZM	toc	*n*		4		4										
		%		50%		100%										
		sign.														
FPT	cross	*n*	4			34										
		%	50%			74%										
		sign.														
HBT	hd	*n*		2		120							4	5		
		%		0%		35%							0%	60%		
		sign.														
LDB	dark	*n*				27			2							
		%				67%			50%							
		sign.														
	light	*n*	4	8		142	3	5	4		3	2		10		
		%	100%	63%		80%	100%	60%	50%		100%	100%		50%		
		sign.														
	trans	*n*	2	7	2	84	2							4		
		%	50%	29%	100%	62%	100%							0%		
		sign.														
NSF	lat	*n*		2		3							3	21		
		%		100%		67%							0%	19%		
		sign.														
OF	cent	*n*				37		trazodone		2		6	2	27		
		%				59%				0%		0%	0%	30%		
		sign.														
	dist	*n*		clonazepam		28				5	3	9	2	56		
		%				18%				0%	0%	44%	0%	13%		
		sign.														
	rear	*n*	2		triazolam	121	2	2		3	3			43		
		%	100%			21%	50%	100%		33%	33%			26%		
		sign.														
	sqrs	*n*	3	2		207	2	2	2	5	6	3	6	69		

		% sign.	33%	0%		25%	100%	100%	50%	20%	17%	33%	0%	17%	
SI	**time**	*n*				14								9	
		% sign.				71%								44%	
STC	**rrs**	*n*			2	19									
		% sign.			50%	79%									
	stps	*n*			2	21									
		% sign.			0%	48%									

Table 4: Number of studies and percentage of positive studies, per combination of test measure and anxiolytic compounds. Cells in grey indicate a percentage of positive studies <75%. Coloured cells highlight a percentage of positive studies >75%. Colour gradient indicates an increasing number of studies. Combinations with only one study were excluded from the table.

DISCUSSION

With the present study, we aimed at providing a synthesis of the reliability of mouse behavioural tests of anxiety. We assessed their sensitivity to a broad range of anxiolytic compounds approved for the treatment of anxiety in humans, using a systematic and unbiased approach. Briefly, we found reported effects to vary greatly across studies and test measures, in addition to overall high heterogeneity and important risks of reporting bias.

We found that for five of the 17 test measures, none of the anxiolytic compounds had a significant effect, whereas, for the remaining 12 test measures, an effect of at least one anxiolytic compound was detected. Additionally, we investigated the overall estimated effect size for each test measure, irrespective of anxiolytic used, and found null or negative overall effects for seven test measures.

For the majority of the test measures and specific compounds, we have observed great variation in the estimated effect sizes, ranging from highly negative to highly positive values, and resulting in estimated cumulative effect sizes close to zero (e.g. in OF-*sqrs* and in OF-*rear*, and in HBT-*hd*.). Additionally, we observed that the effect size estimates of individual studies, which reported a significant effect of a compound also varied greatly even for those combinations in which the overall estimated effect size was positive. Because all of the compounds included in our study were shown to have anxiolytic effects in humans, we consider the proportion of individual studies as a measure of how reliably such behavioural tests can detect behavioural changes elicited by anxiolytic compounds. Overall, only 1254 out of all 2476 contrasts (i.e. 50%) showed significant treatment effects.

Investigation of the total and partial heterogeneity showed that the greater portion (median 74%) of the sample heterogeneity, across test measures, is produced by differences between studies. Such a high level of between-study heterogeneity seems to be common in several fields of animal research (70–73).

There were only two test measures in which the between-study heterogeneity was as low as expected due to random variation alone: SI-*time* and EZM-*toc* These test measures were, however, not sensitive to effects of anxiolytic compounds. Within-study heterogeneity varied greatly across measures but was overall lower than other partial heterogeneity measures, hinting at high levels of standardization within laboratories.

Even though our results show that most of the test measures do not reliably detect behavioural changes elicited by several anxiolytic compounds, we have found two test measures - EPM-*toa* and LDB-*light* - that appear to be sensitive both in terms of detecting a positive effect of anxiolytic compounds and to reliably detect a positive effect in the majority of the individual studies. Additionally, these test measures show significant positive effect sizes for a wider range of anxiolytic compounds than the other measures. With 73% (EPM-*toa*) and 78% (LDB-*light*), respectively, of individual studies reporting a positive effect, the false-negative rates approach the minimally recommended threshold of 0.2. Thus, these measures seem to be promising starting points for refinement and the development of reliable test procedures.

The substantial variation observed between studies using the same test measure and anxiolytic compound with comparable dosages is likely to be attributed to environmental, genetic, and procedural differences. Previous analyses of behavioural test outcomes for the effect of mouse strain on both basal levels of performance and performance after the administration of anxiolytic compounds highlighted substantial strain differences and often conflicting results (46,74–77). Surprisingly, we found only weak effects of mouse strain on heterogeneity for most of the test measures. Apart from genetic background, differences in sex, age, housing conditions, and test environment may contribute to between-study variation. Unfortunately, these are only sporadically and scantily reported. We invite the readers to explore our publicly available dataset through our online application, available at https://mrossovetsuisse.shinyapps.io/Shiny_SR/, which allows displaying data subset by sex, strain, stress treatment and dosage.

Taken together, our results show that most behavioural test measures are unreliable in detecting behavioural changes elicited by anxiolytic compounds other than benzodiazepines and in particular diazepam. This corroborates the previously voiced suspicion that most popular behavioural tests of anxiety are in fact "benzodiazepines tests" (29,47). The behavioural effects elicited by benzodiazepines in these tests have been proposed to reflect disruption of normal behaviour, possibly resulting in altered impulse control rather than attenuated anxiety (47,78).

The behavioural tests included in our study heavily rely on changes in exploration patterns to determine anxiety levels and such test procedures may not be able to disentangle behavioural changes in exploration and anxiety (37,49,79). A clear example of this problem is the open field test, which is sometimes performed to assess anxiety but sometimes to control for locomotor activity in combination with other tests of anxiety (80,81). For example, if the response of animals to a compound is tested in both the LDB and the open field, an increase in LDB-light in the absence of a change in locomotor activity in the open field would suggest that the investigated compound has a specific anxiolytic effect, but no sedative effect, which is highly desirable in anxiolytics especially from a translational perspective (82–84). Upon literature review, we have found as many records in which the open field was performed as a test of locomotor activity (80,81,85,86), as we have found records in which it was performed as a test of anxiety (87–90). Here, we identify an issue with the continuation of such tests as long-held standards that may not be appropriate, due to the researcher's degree of freedom in the interpretation of the test's meaning (91,92).

On a different note, our findings question the standard classification of effect sizes in animal behavioural research. Cohen introduced what are, up to date, considered the conventional thresholds for small, medium, or large effect sizes (namely, a Cohen’s d of 0.2, 0.5, and 0.8 respectively (93)). The author warned for caution (p. 25) in using these thresholds for power analysis outside the scope of the field for which they were initially thought for (psychology or sociology). Study populations of laboratory animals are normally characterised by high

392 degrees of both genetic and environmental standardization (94–96). Therefore, populations of
393 animal studies are usually much more homogenous, producing much lower levels of random
394 variation, when compared to study populations of clinical studies (97). This difference has
395 important implications for the interpretation of standardized effect sizes like Cohen's d or
396 Hedges' g. Due to the higher level of standardization in animal studies and the resulting low
397 within-group variation, a given mean difference between a control and a treatment group will
398 result in a much higher standardized effect size. For example, for EPM-*toa,* (98) reported
399 123.8 seconds spent in the open arms for the control group and 207.3 seconds for the group
400 receiving diazepam. Given the corresponding standard errors of 0.4 and 0.7 for the control
401 and the treatment group, respectively, this amounts to a standardized effect size of 40.6, which
402 is on an entirely different scale of magnitude than a Cohen's d of 0.8, the reference for "large"
403 effects. While this is one of the more extreme examples, we note that EPM-*toa* had an average
404 effect size across drugs of 2.13, with 77% of the total studies reporting an effect size larger
405 than the standard large effect of 0.8. Correct estimation of expected effect sizes is essential
406 for proper power analyses and sample size calculations, with important implications for animal
407 welfare. Considering the large achieved effect sizes, the power analyses based on the
408 "standard Cohen's values" are likely to lead to unnecessarily large required sample sizes.
409 Because of this, we call for a cautious interpretation and more contextualized use of effect
410 size classification, according to each field of research.

411 Our risk of bias assessment showed overall high-risk scores for most of the items. Although
412 the common checklists and tools for risk of bias analyses assess reporting quality rather than
413 study quality, high risks of bias can have serious implications for the reproducibility and
414 replicability of study findings. Albeit efforts have been made to develop more stringent
415 guidelines for both designing and reporting of animal studies (99,100), we observed an overall
416 low quality of reporting, which likely reflects poor study design and conduct. For instance,
417 researchers failed to report the sex or the strain of the animals in 10% of the cases, and
418 important aspects of the housing conditions (e.g. light intensity and temperature),

randomization and blinding procedures, testing conditions (e.g. apparatus size, light intensity, and time of testing), as well as sample size calculations were reported only sporadically.

Our study re-evaluates the suitability of behavioural tests of anxiety in mice, showing low to no sensitivity to anxiolytic compounds (other than diazepam) commonly used for the treatment of anxiety in humans. These finding let us expect poor predictive validity for the discovery of new compounds to treat anxiety disorders in humans and points at a high false-negative rate for individual studies. Additionally, our results highlight considerable idiosyncrasy in the results of the behavioural tests as they are currently performed, with the majority of the tests producing irreproducible and often contradicting results. These findings are corroborated by previous evidence for poor replicability of behavioural tests for anxiety (46,47). Animal tests that lack replicability and validity do not generate new knowledge and, consequently, lose their ethical justification. Additionally, invalid pre-clinical animal trials impair scientific and medical advancement, impacting human subjects in need of treatment. Following the 3Rs principle, effort must be made to improve the quality of animal models for anxiety by developing more informative and reproducible tests with a sound rationale producing results of high internal as well as external validity. This can lead not only to a significant improvement of experimental results but also to more comprehensive and conclusive evidence synthesis in systematic reviews, tackling the prominent bias for positive publications.

Acknowledgements

The authors would like to thank Dr. Cathaljin Leenaars and Dr. Georgia Salanti for their assistance in the data analysis and their valuable feedback on the interpretation of the results.

References

1. Garner JP. The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? ILAR J 2014;55(3):438–56. doi: 10.1093/ilar/ilu047. PubMed PMID: 25541546; PubMed Central PMCID: PMC4342719.
2. Würbel H. More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. Lab Anim (NY) 2017;46(4):164–6. doi: 10.1038/laban.1220. PubMed PMID: 28328898.
3. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med 2013;10(7):e1001489. doi: 10.1371/journal.pmed.1001489. PubMed PMID: 23935460; PubMed Central PMCID: PMC3720257.
4. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature 2014;505(7485):612–3.
5. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res 2015;116(1):116–26. doi: 10.1161/CIRCRESAHA.114.303819. PubMed PMID: 25552691.
6. Roth KA, Cox AE. Science isn't science if it isn't reproducible. Am J Pathol 2015;185(1):2–3. doi: 10.1016/j.ajpath.2014.11.001. PubMed PMID: 25529794.
7. Couzin-Frankel J. When mice mislead. Science 2013;342(6161):922-3, 925. doi: 10.1126/science.342.6161.922. PubMed PMID: 24264972.
8. Perrin S. Make mouse studies work. Nature 2014;507:423–5. doi: 10.1016/j.brainres.2013.10.013. PubMed PMID: 24141148.
9. Russell WMS, Burch RL. The principles of humane experimental technique: Methuen; 1959.
10. Genzel L, Adan R, Berns A, van den Beucken JJJP, Blokland A, Boddeke EHWGM, et al. How the COVID-19 pandemic highlights the necessity of animal research. Current Biology 2020;30(18):R1014-R1018. doi: 10.1016/j.cub.2020.08.030.
11. Directive 2010/63/EU: Additional tools Legislation for the protection of animals used for scientific purposes [Internet].
12. van der Staay FJ, Arndt SS, Nordquist RE. Evaluation of animal models of neurobehavioral disorders. Behav Brain Funct 2009;5:11. doi: 10.1186/1744-9081-5-11. PubMed PMID: 19243583; PubMed Central PMCID: PMC2669803.
13. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? PLoS Med 2010;7(3):e1000245. doi: 10.1371/journal.pmed.1000245. PubMed PMID: 20361020; PubMed Central PMCID: PMC2846855.
14. Willner P. The validity of animal models of depression. Psychopharmacology 1984;83:1–16.
15. Contopoulos-Ioannidis DG, Ntzani EE, Ioannidis JPA. Translation of highly promising basic science research into clinical applications. The American Journal of Medicine 2003;114(6):477–84. doi: 10.1016/S0002-9343(03)00013-5.
16. Geerts H. Of Mice and Men: Bridging the Translational Disconnect in CNS Drug Discovery. CNS Drugs 2009;23(11):915–26.
17. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. Nat Rev Neurol 2014;10(1):37–43. doi: 10.1038/nrneurol.2013.232. PubMed PMID: 24247324.
18. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 2004;3(8):711–5. doi: 10.1038/nrd1470. PubMed PMID: 15286737.

19. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol 2014;32(1):40–51. doi: 10.1038/nbt.2786. PubMed PMID: 24406927.
20. Leenaars CHC, Kouwenaar C, Stafleu FR, Bleich A, Ritskes-Hoitinga M, Vries RBM de, et al. Animal to human translation: a systematic scoping review of reported concordance rates. J Transl Med 2019;17(1):223. doi: 10.1186/s12967-019-1976-2. PubMed PMID: 31307492; PubMed Central PMCID: PMC6631915.
21. Olesen J, Gustavsson A, Svensson M, Wittchen H-U, Jönsson B. The economic cost of brain disorders in Europe. Eur J Neurol 2012;19(1):155–62. doi: 10.1111/j.1468-1331.2011.03590.x. PubMed PMID: 22175760.
22. Kumar V, Bhat ZA, Kumar D. Animal models of anxiety: a comprehensive review. J Pharmacol Toxicol Methods 2013;68(2):175–83. doi: 10.1016/j.vascn.2013.05.003. PubMed PMID: 23684951.
23. Harro J. Animals, anxiety, and anxiety disorders: How to measure anxiety in rodents and why. Behavioural Brain Research 2018;352:81–93. doi: 10.1016/j.bbr.2017.10.016. PubMed PMID: 29050798.
24. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet 2016;388(10053):1545–602. doi: 10.1016/S0140-6736(16)31678-6.
25. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 2005;62:593–602.
26. Ravindran LN, Stein MB. The Pharmacologic Treatment of Anxiety Disorders: A Review of Progress. The Journal of clinical psychiatry 2010;71(7):839–54.
27. Crawley JN. What's wrong with my mouse?: Behavioral phenotyping of transgenic and knockout mice. 2nd ed. Hoboken N.J.: Wiley-Interscience; 2007. xvi, 523.
28. Belzung C, Griebel G. Measuring normal and pathological anxiety-like behaviour in mice: a review. Behavioural Brain Research 2001;125(1-2):141–9. doi: 10.1016/S0166-4328(01)00291-1.
29. Ennaceur A. Tests of unconditioned anxiety - pitfalls and disappointments. Physiol Behav 2014;135:55–71. doi: 10.1016/j.physbeh.2014.05.032. PubMed PMID: 24910138.
30. Hånell A, Marklund N. Structured evaluation of rodent behavioral tests used in drug discovery research. Front Behav Neurosci 2014;8:252. doi: 10.3389/fnbeh.2014.00252. PubMed PMID: 25100962; PubMed Central PMCID: PMC4106406.
31. Walsh RN, Cummins RA. The Open-Field Test: A Critical Review. Psychological Bulletin 1976;83(3):482–504.
32. Belzung C. Measuring rodent exploratory behavior. Handbook of Molecular-Genetic Techniques for Brain and Behavior Research 1999;13.
33. Hall CS. Emotional behavior in the rat. I. Defecation and urination as measures of individual differences in emotionality. Journal of Comparative Psychology 1934;18(3):385–403.
34. DeFries JC, Hegmann JP, Halcomb RA. Response to 20 Generations of Selection for Open-Field Activity in Mice. Behavioural Biology 1974;11:481–95.
35. DeFries JC, Hegmann JP, Weir, Morton, W. Open-Field Behavior in Mice: Evidence-for a Major Gene Effect Mediated by the Visual System. Science 1966;154(3756):1577–9.
36. Prut L, Belzung C. The open field as a paradigm to measure the effects of drugs on anxiety-like behaviors: a review. European Journal of Pharmacology 2003;463(1-3):3–33. doi: 10.1016/S0014-2999(03)01272-X.

37. Bourin M, Petit-Demoulière B, Dhonnchadha BN, Hascöet M. Animal models of anxiety in mice. Fundam Clin Pharmacol 2007;21(6):567–74. doi: 10.1111/j.1472-8206.2007.00526.x. PubMed PMID: 18034657.
38. Montgomery KC. The relation between fear induced by novel stimulation and exploratory behavior. J. Comp. Physiol. PsychoL 1958;48:254–60.
39. Crawley J, Goodwin FK. Preliminary Report of a Simple Animal Behavior Model for the Anxiolytic Effects of Benzodiazepines. Pharmacology Biochemistry and Behavior 1980;13:167–70.
40. Ohl F, Arndt SS, van der Staay FJ. Pathological anxiety in animals. Vet J 2008;175(1):18–26. doi: 10.1016/j.tvjl.2006.12.013. PubMed PMID: 17321766.
41. File SE, Wardill AG. The reliability of the hole-board apparatus. Psychopharmacologia 1975;44(1):47–51. doi: 10.1007/BF00421183.
42. Shepherd JK, Grewal SS, Fletcher A, Bill DJ, Dourish CT. Behavioural and pharmacological characterisation of the elevated "zero-maze" as an animal model of anxiety. Psychopharmacology 1994;116(1):56–64. doi: 10.1007/BF02244871.
43. File SE, Hyde JRG. A Test of Anxiety that Distinguishes Between the Actions of Benzodiazepines and Those of Other Minor Tranquilisers and of Stimulants. Pharmacology Biochemistry and Behavior 1979;11:65–9.
44. Bodnoff SR, Suranyi-Dacotte B, Aitken, David H., Quirion, Remi, Meaney, Michael. The effects of chronic antidepressant treatment in an animal model of anxiety. Psychopharmacology 1988;95:298–302.
45. Aron C, Simon P, Larousse C, Boissier JR. Evaluation of a rapid technique for detecting minor tranquilizers. Neuropharmacology 1971;10(4):459–69. doi: 10.1016/0028-3908(71)90074-8.
46. Ennaceur A, Chazot PL. Preclinical animal anxiety research - flaws and prejudices. Pharmacol Res Perspect 2016;4(2):e00223. doi: 10.1002/prp2.223. PubMed PMID: 27069634; PubMed Central PMCID: PMC4804324.
47. Bespalov A, Steckler T. Pharmacology of Anxiety or Pharmacology of Elevated Plus Maze? Biological psychiatry 2021. doi: 10.1016/j.biopsych.2020.11.026. PubMed PMID: 33612186.
48. Cryan JF, Sweeney FF. The age of anxiety: role of animal models of anxiolytic action in drug discovery. Br J Pharmacol 2011;164(4):1129–61. doi: 10.1111/j.1476-5381.2011.01362.x. PubMed PMID: 21545412; PubMed Central PMCID: PMC3229755.
49. Hascoët M, Bourin M. A New Approach to the Light/Dark Test Procedure in Mice. Pharmacology Biochemistry and Behavior 1998;60(3):645–53. doi: 10.1016/S0091-3057(98)00031-8.
50. Rodgers RJ, Cao B-J, Dalvi A, Holmes A. Animal models of anxiety: an ethological perspective. Braz J Med Biol Res 1997;30(3):289–304. doi: 10.1590/S0100-879X1997000300002.
51. Bystrisky A, Khalsa SS, Cameron MR, Schiffman J. Current Diagnosis and Treatment of Anxiety Disorders. Pharmacy & Therapeutics 2013;38(1):30–44.
52. Costa JP, Oliveira GAL de, Almeida AAC de, Islam MT, Sousa DP de, Freitas RM de. Anxiolytic-like effects of phytol: possible involvement of GABAergic transmission. Brain Res 2014;1547:34–42. doi: 10.1016/j.brainres.2013.12.003. PubMed PMID: 24333358.
53. Moniruzzaman M, Mannan MA, Hossen Khan MF, Abir AB, Afroze M. The leaves of Crataeva nurvala Buch-Ham. modulate locomotor and anxiety behaviors possibly through GABAergic system. BMC Complement Altern Med 2018;18(1):283. doi: 10.1186/s12906-018-2338-y. PubMed PMID: 30340574; PubMed Central PMCID: PMC6194725.
54. Borsini F, Podhorna J, Marazziti D. Do animal models of anxiety predict anxiolytic-like effects of antidepressants? Psychopharmacology 2002;163(2):121–41. doi: 10.1007/s00213-002-1155-6. PubMed PMID: 12202959.

55. Hart BA 't. Reverse translation of failed treatments can help improving the validity of preclinical animal models. European Journal of Pharmacology 2015;759:14–8. doi: 10.1016/j.ejphar.2015.03.030. PubMed PMID: 25823810.
56. Shakhnovich V. It's Time to Reverse our Thinking: The Reverse Translation Research Paradigm. Clin Transl Sci 2018;11(2):98–9. doi: 10.1111/cts.12538. PubMed PMID: 29423973; PubMed Central PMCID: PMC5866972.
57. Baker M. Reproducibility crisis: Blame it on the antibodies. Nature 2015;521(7552):274–6.
58. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLoS Biol 2015;13(6):e1002165. doi: 10.1371/journal.pbio.1002165. PubMed PMID: 26057340; PubMed Central PMCID: PMC4461318.
59. Baker M. Is there a Reproducibility Crisis? 1,500 scientists lift the lid on reproducibility. Nature 2016;533(7604):452-454. doi: 10.1126/science.aac4716. PubMed PMID: 26315443.
60. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain 2020;13(1):24. doi: 10.1186/s13041-020-0552-2. PubMed PMID: 32079532.
61. Smith AJ, Lilley E. The Role of the Three Rs in Improving the Planning and Reproducibility of Animal Experiments. Animals (Basel) 2019;9(11). doi: 10.3390/ani9110975. PubMed PMID: 31739641; PubMed Central PMCID: PMC6912437.
62. Hackam DG, Redelmeier DA. Translation of Research Evidence From Animals to Humans. The Journal of the American Medical Association 2006;296(14):1731–2.
63. Mak IWY, Evaniew N, Ghert M. Lost in translation: animal models and clinical trials in cancer treatment. American Journal of Translational Research 2014;6(2):114–8.
64. O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 1,026 experimental treatments in acute stroke. Ann Neurol 2006;59(3):467–77. doi: 10.1002/ana.20741. PubMed PMID: 16453316.
65. Garner JP, Gaskill BN, Weber EM, Ahloy-Dallaire J, Pritchett-Corning KR. Introducing Therioepistemology: the study of how knowledge is gained from animal research. Lab Anim (NY) 2017;46(4):103–13. doi: 10.1038/laban.1224. PubMed PMID: 28328885.
66. Steimer T. Animal models of anxiety disorders in rats and mice: some conceptual issues. Dialogues in clinical neurosciences 2011;13(4):495–506.
67. R Core Team. R: A Language and Environment for Statistical Computing: Vienna, Austria; 2020.
68. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software 2010;36(3).
69. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. Res Synth Methods 2011;2(1):61–76. doi: 10.1002/jrsm.35. PubMed PMID: 26061600.
70. Pires GN, Bezerra AG, Tufik S, Andersen ML. Effects of experimental sleep deprivation on anxiety-like behavior in animal research: Systematic review and meta-analysis. Neurosci Biobehav Rev 2016;68:575–89. doi: 10.1016/j.neubiorev.2016.06.028. PubMed PMID: 27345144.
71. Antoniuk S, Bijata M, Ponimaskin E, Wlodarczyk J. Chronic unpredictable mild stress for modeling depression in rodents: Meta-analysis of model reliability. Neurosci Biobehav Rev 2019;99:101–16. doi: 10.1016/j.neubiorev.2018.12.002. PubMed PMID: 30529362.
72. Leffa DT, Panzenhagen AC, Salvi AA, Bau CHD, Pires GN, Torres ILS, et al. Systematic review and meta-analysis of the behavioral effects of methylphenidate in the spontaneously hypertensive rat model of attention-deficit/hyperactivity disorder. Neurosci Biobehav Rev 2019;100:166–79. doi: 10.1016/j.neubiorev.2019.02.019. PubMed PMID: 30826386.
73. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol 2018;16(2):e2003693.

doi: 10.1371/journal.pbio.2003693. PubMed PMID: 29470495; PubMed Central PMCID: PMC5823461.

74. Griebel G, Belzung C, Perrault G, Sanger DJ. Differences in anxiety-related behaviours and in sensitivity to diazepam in inbred and outbred strains of mice. Psychopharmacology 2000;148:164–70.
75. Bouwknecht JA, Paylor R. Behavioral and physiological mouse assays for anxiety: a survey in nine mouse strains. Behavioural Brain Research 2002;136(2):489–501. doi: 10.1016/S0166-4328(02)00200-0.
76. Hagenbuch N, Feldon J, Yee BK. Use of the elevated plus-maze test with opaque or transparent walls in the detection of mouse strain differences and the anxiolytic effects of diazepam. Behavioural Pharmacology 2006;17:31–41.
77. Gard PR, Haigh SJ, Cambursano PT, Warrington CA. Strain differences in the anxiolytic effects of losartan in the mouse. Pharmacology Biochemistry and Behavior 2001;69:35–40.
78. Thiébot M-H, Soubrié P, Simon P. Is delay of reward mediated by shock-avoidance behavior a critical targer for anti-punishment effects of diazepam in rats? Psychopharmacology 1985;87:473–9.
79. Andreatini R, Bacellar LFS. Animal models: Trait or state measure? The test-retest reliability of the elevated plus-maze and behavioral despair. Progress in Neuro-Psychopharmacology and Biological Psychiatry 2000;24(4):549–60. doi: 10.1016/S0278-5846(00)00092-0.
80. Devi MR, Bawari M, Paul SB. Neurotoxic effect of Albizia myriophylla benth., a medicinal plant in male mice. International Journal of Pharmacy and Pharmaceutical Sciences 2013;5(3).
81. Akbar S, Subhan F, Karim N, Aman U, Ullah S, Shahid M, et al. Characterization of 6-methoxyflavanone as a novel anxiolytic agent: A behavioral and pharmacokinetic approach. European Journal of Pharmacology 2017;801:19–27. doi: 10.1016/j.ejphar.2017.02.047. PubMed PMID: 28257822.
82. Grundmann O, Nakajima J-I, Seo S, Butterweck V. Anti-anxiety effects of Apocynum venetum L. in the elevated plus maze test. J Ethnopharmacol 2007;110(3):406–11. doi: 10.1016/j.jep.2006.09.035. PubMed PMID: 17101250.
83. Lolli LF, Sato CM, Romanini CV, Villas-Boas LDB, Santos CAM, Oliveira RMW de. Possible involvement of GABA A-benzodiazepine receptor in the anxiolytic-like effect induced by Passiflora actinia extracts in mice. J Ethnopharmacol 2007;111(2):308–14. doi: 10.1016/j.jep.2006.11.021. PubMed PMID: 17196350.
84. Tabari MA, Tehrani MAB. Evidence for the involvement of the GABAergic, but not serotonergic transmission in the anxiolytic-like effect of bisabolol in the mouse elevated plus maze. Naunyn Schmiedebergs Arch Pharmacol 2017;390(10):1041–6. doi: 10.1007/s00210-017-1405-0. PubMed PMID: 28730280.
85. Adeoluwa OA, Aderibigbe AO, Agu GO, Adewole FA, Eduviere AT. Neurobehavioural and Analgesic Properties of Ethanol Bark Extract of Terminalia ivorensis A Chev. (Combrataceae) in Mice. Drug Res (Stuttg) 2015;65(10):545–51. doi: 10.1055/s-0034-1394417. PubMed PMID: 25514116.
86. Moreira DRM, Santos DS, Espírito Santo RFd, Santos FED, Oliveira Filho GB de, Leite ACL, et al. Structural improvement of new thiazolidinones compounds with antinociceptive activity in experimental chemotherapy-induced painful neuropathy. Chem Biol Drug Des 2017;90(2):297–307. doi: 10.1111/cbdd.12951. PubMed PMID: 28112878.
87. Ang HH, Cheang HS. Studies on the Anxiolytic Activity of Eurycome longifolia Jack Roots in Mice. Japanese Journal of Pharmacology 1999;76:497–500.
88. Sonovane GS, Sarveiya VP, Kasture VS, Kasture SB. Anxiogenic activity of Myristica fragrans seeds. Pharmacology Biochemistry and Behavior 2002;71:239–44.

89. Figueredo YN, Rodríguez EO, Reyes YV, Domínguez CC, Parra AL, Sánchez JR, et al. Characterization of the anxiolytic and sedative profile of JM-20: a novel benzodiazepine-dihydropyridine hybrid molecule. Neurol Res 2013;35(8):804–12. doi: 10.1179/1743132813Y.0000000216. PubMed PMID: 23651620.
90. Ketcha Wanda GJM, Djiogue S, Gamo FZ, Ngitedem SG, Njamen D. Anxiolytic and sedative activities of aqueous leaf extract of Dichrocephala integrifolia (Asteraceae) in mice. J Ethnopharmacol 2015;176:494–8. doi: 10.1016/j.jep.2015.11.035. PubMed PMID: 26602454.
91. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Front Psychol 2016;7:1832. doi: 10.3389/fpsyg.2016.01832. PubMed PMID: 27933012; PubMed Central PMCID: PMC5122713.
92. Pound P, Bracken MB. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? BMJ 2014;348:g3387. doi: 10.1136/bmj.g3387. PubMed PMID: 24879816.
93. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Academic Press; 1977.
94. Wahlsten D, Rustay NR, Metten P, Crabbe JC. In search of a better mouse test. Trends in Neurosciences 2003;26(3):132–6. doi: 10.1016/S0166-2236(03)00033-X.
95. Wahlsten D. Standardizing tests of mouse behavior: Reasons, recommendations, and reality. Physiol Behav 2001;73(5):695–704. doi: 10.1016/S0031-9384(01)00527-3.
96. Würbel H. Behaviour and the standardization fallacy. Nature genetics 2000;26:262–3.
97. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci 2020;21(7):384–93. doi: 10.1038/s41583-020-0313-3. PubMed PMID: 32488205.
98. Santana LCLR, Brito MRM, Oliveira GLS, Citó AMGL, Alves CQ, David JP, et al. Mikania glomerata: Phytochemical, Pharmacological, and Neurochemical Study. Evid Based Complement Alternat Med 2014;2014:710410. doi: 10.1155/2014/710410. PubMed PMID: 25202336; PubMed Central PMCID: PMC4151546.
99. Du Percie Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLoS Biol 2020;18(7):e3000410. doi: 10.1371/journal.pbio.3000410. PubMed PMID: 32663219; PubMed Central PMCID: PMC7360023.
100. Smith AJ, Clutton RE, Lilley E, Hansen KEA, Brattelid T. PREPARE: guidelines for planning animal research and testing. Lab Anim 2018;52(2):135–41. doi: 10.1177/0023677217724823. PubMed PMID: 28771074; PubMed Central PMCID: PMC5862319.

Fig 1 - **Flowchart of the screened papers and reasons for exclusion**. ss: unclear or absent sample size; unav: paper unavailable; par: incompatible outcomes reported; drug: incompatible compounds used; par-report: issues with the reporting of the outcomes; paper: wrong format of paper; test: incompatible behavioural test used; animal: wrong animals used; sem-sd: unclear or absent measure of variance; other.

Fig 2: Violin plots showing the probability density distribution of the calculated effect size (x-axis) of the individual studies for each test measure. Overlapped to the violin plots, the overall estimated effect size for each test measure, indicated by the diamonds, and the relative 95% confidence interval. Points indicate the estimated mean effect size for each compound. Colours indicate anxiolytic compounds. Opacity is applied to not significant effect sizes, i.e. the lower bound of the 95% confidence interval is lower than zero. An interactive version of the Fig can be found online at https://mrossovetsuisse.shinyapps.io/Shiny_SR/.

Fig 3: **Forest plots of three selected test measures**: A: LDB-*light*, B: EPM-*toa*, C: OF-*sqrs*, sorted for increasing effect size. Different colours indicate different anxiolytic compounds, as indicated in the legend (See Supplementary material for remaining measures).

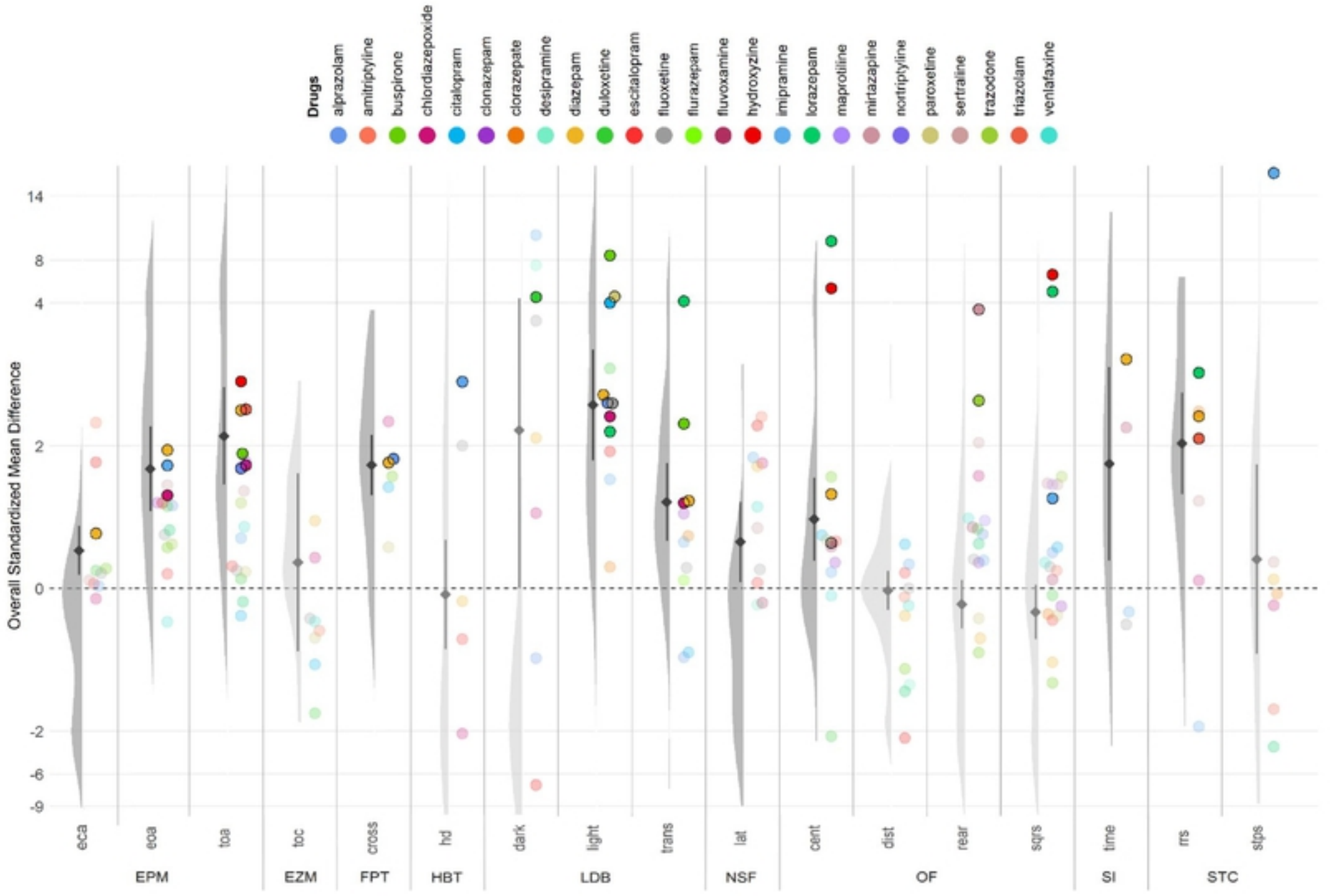
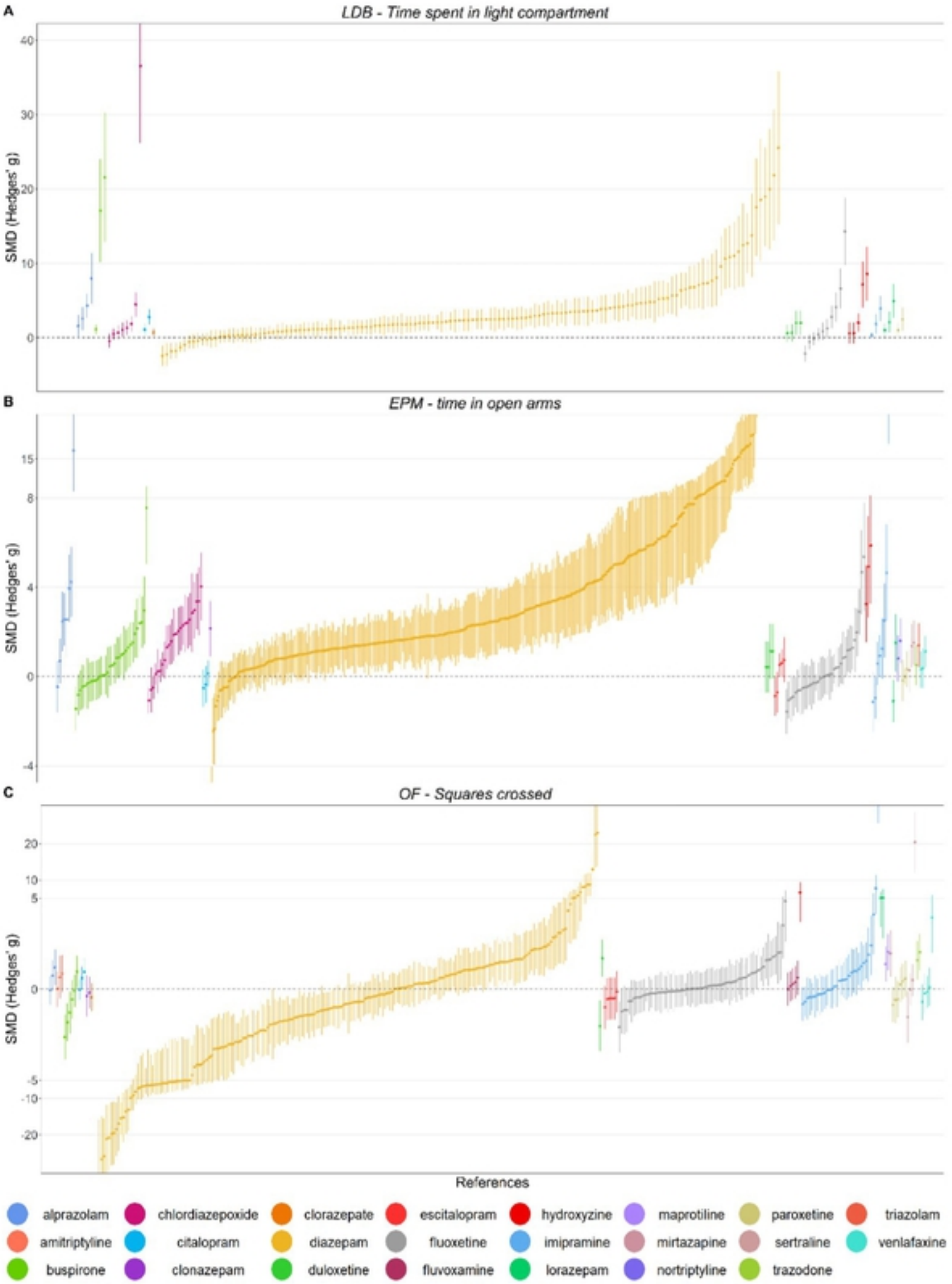

Drugs
alprazolam
amitriptyline
buspirone
chlordiazepoxide
citalopram
clonazepam
clorazepate
desipramine
diazepam
duloxetine
escitalopram
fluoxetine
flurazepam
fluvoxamine
hydroxyzine
imipramine
lorazepam
maprotiline
mirtazapine
nortriptyline
paroxetine
sertraline
trazodone
triazolam
venlafaxine
Overall Standardized Mean Difference
14
8
4
2
0
-2
-6
-9
eca
eoa
toa
toc
cross
hd
dark
light
trans
lat
cent
dist
rear
sqrs
time
rrs
stps
EPM
EZM
FPT
HBT
LDB
NSF
OF
SI
STC

Figure 2

Figure 3


A
LDB - Time spent in light compartment
SMD (Hedges' g)
40
30
20
10
0
B
EPM - time in open arms
SMD (Hedges' g)
15
8
4
0
-4
C
OF - Squares crossed
SMD (Hedges' g)
20
10
5
0
-5
-10
-20
References
alprazolam
amitriptyline
buspirone
chlordiazepoxide
citalopram
clonazepam
clorazepate
diazepam
duloxetine
escitalopram
fluoxetine
fluvoxamine
hydroxyzine
imipramine
lorazepam
maprotiline
mirtazapine
nortriptyline
paroxetine
sertraline
trazodone
triazolam
venlafaxine

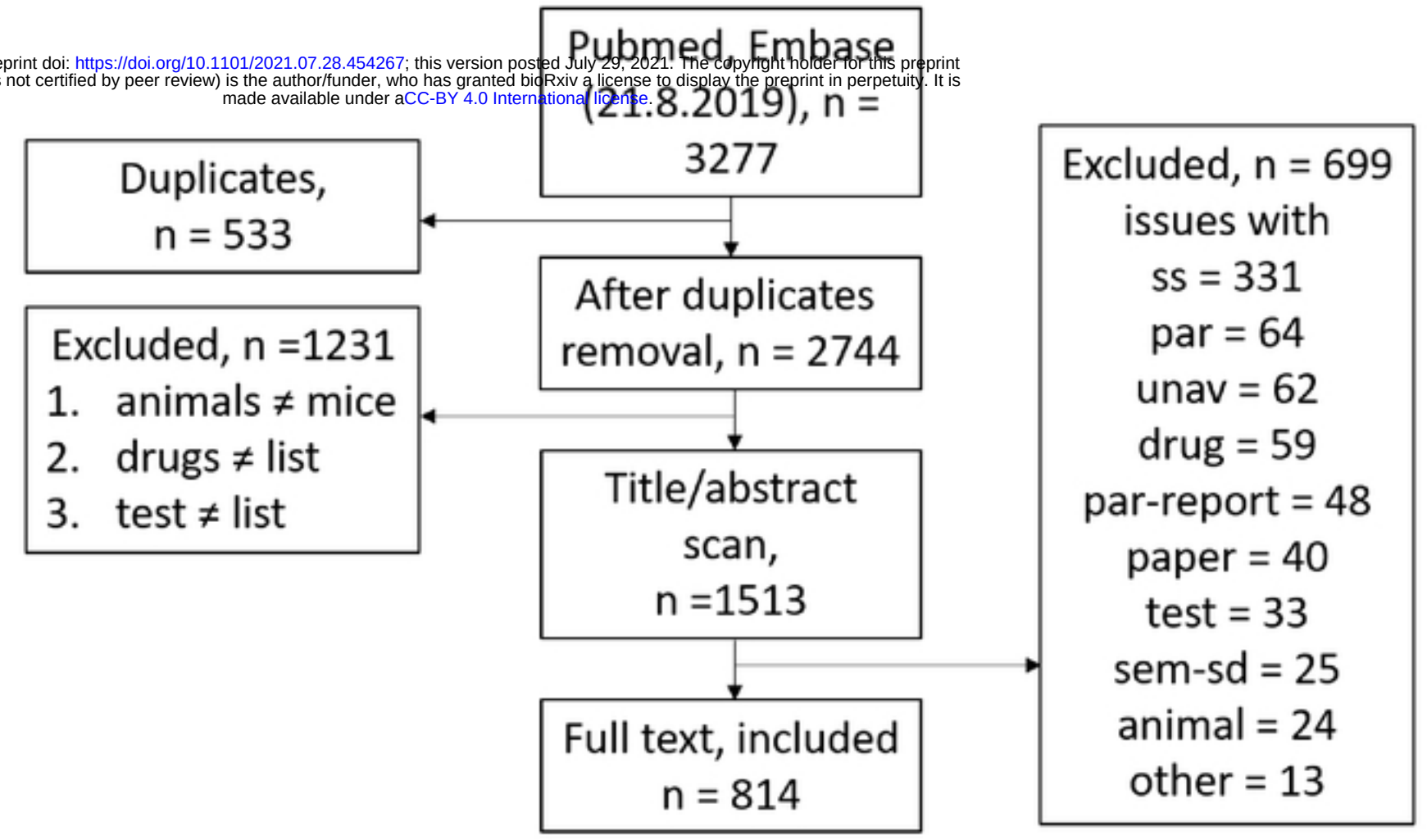


Figure 1