The neurophysiological potential

Kat, Renate

DOI:
10.33612/diss.249073151

Publication date:
2022

Citation for published version (APA):
Kat, R. (2022). *The neurophysiological potential: Translatability and reproducibility of preclinical outcome measures in psychiatric research*. University of Groningen. https://doi.org/10.33612/diss.249073151

CHAPTER 3

Translational validity and methodological underreporting in animal research: a systematic review and meta-analysis of the Fragile X syndrome (Fmr1 KO) rodent model

Renate Kat[a,c], María Arroyo-Araujo[a,c], Rob B.M. de Vries[b], Marthe A. Koopmans[a], Sietse F. de Boer[a], Martien J.H. Kas[a]

[a] *Groningen Institute for Evolutionary Life Sciences, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands*

[b] *SYRCLE, Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Centre, Geert Groteplein Zuid 21, 6525 EZ, Nijmegen, The Netherlands,*

[c] *These authors contributed equally to this work*

Published in Neuroscience and Biobehavioural Reviews (2022, 139:104722)

Abstract

Predictive models are essential for advancing knowledge of brain disorders. High variation in study outcomes hampers progress. To address the validity of predictive models, we performed a systematic review and meta-analysis on behavioural phenotypes of the knock-out rodent model for Fragile X syndrome according to the PRISMA reporting guidelines. In addition, factors accountable for the heterogeneity between findings were analyzed. The knock-out model showed good translational validity and replicability for hyperactivity, cognitive and seizure phenotypes. Despite low replicability, translational validity was also found for social behaviour and sensory sensitivity, but not for attention, aggression and cognitive flexibility. Anxiety, acoustic startle and prepulse inhibition phenotypes, despite low replicability, were opposite to patient symptomatology. Subgroup analyses for experimental factors moderately explain the low replicability, these analyses were hindered by under-reporting of methodologies and environmental conditions. Together, the model has translational validity for most clinical phenotypes, but caution must be taken due to low effect sizes and high inter-study variability. These findings should be considered in view of other rodent models in preclinical research.

Keywords: *Autism spectrum disorder, mouse models, preclinical data quality*

Introduction

The Fragile X Syndrome (FXS) is a common inherited form of intellectual disability and one of the most prominent genetic causes of syndromic autistic spectrum disorders (ASD; Kidd et al., 2014). FXS is caused by a CGG repeat mutation on the X chromosome containing the *FMR1* gene, causing a deficiency of the resultant protein (Verkerk et al., 1991). The *FMR1* gene codes for the RNA-binding protein fragile X mental retardation protein (FMRP), which binding targets include several synaptic proteins essential for proper neurotransmission and neuronal structure, affecting multiple neuronal pathways. Individuals carrying the full FMRP mutation typically display intellectual disabilities, seizures, attention deficits, increased anxiety, hyperarousal to stimuli, and macroorchidism together with autistic-like features. Of the FXS patients, 30% meet the criteria for ASD diagnosis (Bailey et al., 1998; Baumgardner et al., 1995; Hagerman et al., 1986; Hersh et al., 2011), but up to 90% of patients show some of the symptoms of ASD (Hagerman et al., 1986). In general, females display milder symptoms than males.

The FMRP lack of expression was successfully reproduced in mice to generate an animal model to study. The most frequently used Fmr1 KO mouse model came from the Dutch-Belgian Fragile X Consortium (The Dutch-Belgian Fragile X Consortium et al., 1994) and does not produce FMRP because of a disruption in the *FMR1* DNA sequence with an insertion in exon 5. Still, it has a detectable level of *FMR1* mRNA (Kazdoba et al., 2014). A second-generation KO model (KO2) was later developed which no longer has *Fmr1* mRNA present (Mientjes et al., 2006). The majority of research on these models have focussed on the affected molecular pathophysiological pathways, like increased immature spine densities and GABA-ergic deficits, which have recently been reviewed elsewhere (Dionne and Corbin, 2021; Telias, 2019). Additionally, a large body of literature has reported on the behavioural abnormalities of these models. These mouse models, as well as some KO rat models, have been reported to recapitulate several phenotypic features seen in patients such as cognitive deficits, social anxiety, reduced social interaction, repetitive behaviours and hyperactivity. However, there is a considerable number of contrasting findings in the literature. For example, while many papers report inhibitory avoidance cognitive deficits in *Fmr1* KO mice (Ding et al., 2020, 2014; Li et al., 2020; Qin et al., 2015; Saré et al., 2016) other studies found no difference between KO and wildtype (WT) mice using the same task (Liao et al., 2018; Melancia and Trezza, 2018; Saré et al., 2019, 2018; The Dutch-Belgian Fragile X Consortium et al., 1994). These discrepancies are also found for tasks that test for recognition memory, social discrimination, and spatial memory and, more importantly, for tasks that measure the core symptomatic features of ASD such as tasks that evaluate social behaviour, repetitive behaviour, communication, and anxiety. Recently in our lab, the mouse *Fmr1* KO model was tested in a behavioural battery to assess repetitive and social behaviours. To our surprise, no apparent phenotype was found, also contrasting with the behavioural repertoire seen in patients and sometimes found in the preclinical literature.

It has been suggested that differences in methodological approaches and diverse research practices can impact the behavioural outcome measures, which may partly explain the contrasting literature. However, preclinical research has also shown a lack of transparency of reporting as well as the use of inappropriate statistical analysis and insufficient sample sizes (Kilkenny et al., 2009; Prinz et al., 2011) putting the validity, replicability and translatability of results at stake.

The divergent results of the *Fmr1* KO phenotype raise questions about the validity as a preclinical model of neurodevelopmental disorders. In general, molecular studies quantifying the null expression of FMRP and its consequences on molecular alterations reach consensus. However, behavioural studies tend to show more discrepancies across laboratories and tasks. These discrepancies could suggest that the way in which the phenotype is assessed is not appropriate; for example, poorly sensitive tasks or poor experimental design, both of which are relevant for the internal and face validity of the model (Belzung and Lemoine, 2011; Campbell and Stanley, 1963). Additionally, it may be that the phenotype is not robust enough and therefore it only shows in some scenarios but can't be generalized to other study samples and/or scenarios, which questions the model's external validity (Campbell and Stanley, 1963; Richter, 2017). In order to objectively evaluate the phenotype of the *Fmr1* KO it is necessary to review the available literature and evaluate its methods.

Systematic review and meta-analysis are valuable tools to make a transparent (statistical) summary of research findings that yields an estimate of the validity of the overall findings. Although their use in preclinical science is relatively new, their value has been appraised by various disciplines such as medical sciences, psychology and education. By looking at the range of available published studies one can judge the external validity in addition to the possibility of assessing the risk of a publication bias. On the other hand, an indication of the internal validity based on a risk of bias assessment informs us about the methodological quality of the included studies overall (Sena et al., 2014). In this way, the systematic review and meta-analysis presented here will shed some light on the behavioural phenotype of the *Fmr1* KO line. Additionally, it will give insight into the experimental factors which affect the genotype expression and thereby potentially contribute to the variability in results presented in literature. In addition, an indication of the reporting quality in the field and a publication bias will be discussed further to properly ponder the results.

Given the large amount of available behavioural studies available in the literature, we decided to narrow down our systematic review and meta-analysis to the behavioural categories that are most relevant to evaluate the FXS/ASD-like phenotype. In the case of autism-like behaviours, we chose to focus on social behaviours, repetitive behaviours, anxiety, sensory gating, and sensory sensitivity as these are often reported in patients. In addition, learning, memory, and attention performance are relevant for the model given the intellectual disability component of FXS and thus, the *Fmr1* model (Harris et al., 2008). Locomotion was chosen based on its wide use given that most genetically modified models exhibit hyperactivity. Lastly, audiogenic seizures have high comorbidity with epilepsy as well as ASD and its increased excitation/inhibition (E/I) balance hypothesis.

Methods

The review protocol was preregistered on PROSPERO (www.crd.york.ac.uk/prospero; CRD42020191070). The reporting in this systematic review adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Page et al., 2021; Supplementary file 12).

Search strategy

Two bibliographic databases were systematically searched for relevant studies: Pubmed and Web of Science. The search consisted of two components, one for fragile X syndrome and *Fmr1*, and one for rat and mouse (for full strategy see Supplementary file 1). If available, both controlled terms (*i.e.*, MeSH), and free text words were used. Bibliographic results were imported and de-duplicated using Rayyan software (Ouzzani et al., 2016). The final search was performed on 30-06-2021. In addition, the reference lists of all included studies were scanned for relevant studies that did not come up in the bibliographic search.

Eligibility screening

Studies were eligible for inclusion if they compared wildtype (WT) and knockout (KO) rodents for *Fmr1* in one or more behavioural tests relevant for our domains of interest (Supplementary file 2). The domains of interest included: locomotion, social behaviour (sociability, aggression, communication and social cognition), learning and cognition (conditioned learning, spatial learning, recognition learning and working memory), repetitive behaviour (low order repetitive behaviour and cognitive flexibility), anxiety, attention, sensory sensitivity (olfactory, somatosensory, auditory, visual and nociception) and sensitivity for audiogenic seizures. MA and RK independently screened all identified records in two stages using Rayyan software. Disagreements were resolved by discussion.

The first stage concerned screening of the title and abstract of the articles. In this stage articles were excluded for the following reasons: (i) not an original primary study (*e.g.*, review, editorial or conference abstract), (ii) the used model was not a mouse or rat, (iii) the used model was not an *Fmr1* knockout (CGG-repeat knock-in models were not included), (iv) in vitro or ex vivo studies where behavioural assessment is impossible.

In the second stage, the full text of the remaining articles was screened. Articles were excluded for one or more of the reasons from stage one, plus the following additional reasons: (v) not a full KO (*e.g.*, selective or conditional KOs) or no use of WT control, (vi) no control condition for additional interventions present (*e.g.*, vehicle), (vii) behavioural tasks that did not fit with the behavioural domains of interest, (viii) no full text available.

Extraction of study characteristics

Extraction of study characteristics was performed by MA and RK, who both extracted characteristics for half of the studies. MA, a native speaker of the Spanish language, extracted the article written in Spanish. The following study characteristics were extracted: (i) study ID: first author, last author, year, journal, digital object identifier (DOI), article language; (ii) animal model characteristics: species, genetic background, sex, age, KO or KO2 (for mouse studies), being littermates; (iii) study design characteristics: housing conditions (group housed - mixed genotypes; group-housed - same genotypes; single housed), presence of additional interventions, test phase (light or dark), number of behavioural tasks; (iv) outcome measures: list of (relevant) behavioural tests used, list of test outcomes used.

Risk of bias assessment

Risk of bias assessment was performed to assess the methodological quality of the studies included in the meta-analysis. Due to the high number of papers included in the current study, and the high percentage of ‘unclear risk of bias’ scores expected because of poor reporting, the risk of bias analysis was performed on a random sample of 45 papers (18%). The SYRCLE risk of bias tool for animal studies (Hooijmans et al., 2014) was used. Item one and three from the tool, concerning randomization and blinding of treatment allocation, were only assessed for studies performing additional pharmacological interventions of which the vehicle groups were used in the current meta-analysis. Item seven (random outcome assessment) was scored as low risk of bias when data was scored using computerized automated scoring. Items were answered with a “Yes” for low risk of bias, “No” for high risk of bias and “Unclear” if it was not possible to assess the risk of bias due to lack of information. Risk of bias assessment was independently performed by MA and RK, disagreements were resolved by discussion.

Extraction of outcome data

For every study data was extracted for each behavioural domain in which thebehavioural tests were performed. Mean, standard deviation (SD) or standard error (SEM) and the number of animals (N) were extracted for the WT and KO groups. For audiogenic seizures, the number of animals that did or did not experience seizures and the sample size was extracted for both WT and KO groups. If percentages of animals experiencing seizures were reported, the number of animals was calculated using the total sample size. Whenever possible, exact values were taken from text or tables. When those were not available, WebPlotDigitizer software (v3.8-4.4, Rohatgi, A., Pacifica, CA, USA, https://automeris.io/WebPlotDigitizer) was used to extract the numbers from figures. Although the initial protocol stated that authors would be contacted when using WebPlotDigitizer was not possible, we decided to refrain from this, due to the high number of studies and amount of data already included in the study. When it was unclear whether SD or SEM was reported, SEM was assumed, in order to be more conservative. Data extraction was performed by MA, RK and MAK. A random sample of studies (11, 4.3%) was extracted twice at the start to check referees' reliability. When ranges were reported for N, the highest value of the range was used to calculate the SD in case the study reported the SEM (SD = SEM*√N), while the lowest value of the range was used as sample size in the actual meta-analysis (Ramsteijn et al., 2020). If a group of animals was used in comparison to multiple other groups (*e.g.*, WT females compared to both heterozygous and homozygous KO females), an adjusted sample size was used in the meta-analysis (sample size divided by the number of comparisons in which this group is used).

Before starting the extraction of outcome data, a categorization and prioritization of behavioural tests and outcomes was made by MA, RK and MAK and later discussed with MJHK and SB. All behavioural tests used within the included studies were allocated to one of the (sub-) domains of interest (Supplementary file 2). Within every (sub-)domain behavioural tests were ranked from most to least relevant, and for every test, outcome measures were ranked from most to least relevant. This ranking guided the data extraction, to assure that in the case of multiple reported outcomes or even multiple reported tests within the same behavioural domain, unique animals appeared only once in every domain. If a study performed experiments in multiple groups of animals (*e.g.*, males and females, different age groups or multiple additional interventions) we analysed these comparisons as if they were separate studies.

For social tasks, although social malfunctioning may be also expressed in male-female socio-sexual interactions and male-juvenile explorations across different ages, we decided to prioritize adult male-male interactions to characterise an adult phenotype independent of sexual and neurodevelopmental maturity. Furthermore, adult male-male interactions are the most frequently used for social interaction paradigms in studies on *Fmr1* KO mice (*i.e.*, 45% of all reported social interaction, against 15% for male-juvenile, 10% for male-female and 8% for female-female). Ultrasonic vocalisation (USV) data was pooled over all call-types and frequencies. For cognitive tasks, including cognitive flexibility, the data from the last trial was always used to assess a stable outcome measure independent of the learning process. For recognition learning, the test with retention time closest to one hour was used. In the 5-choice serial reaction time task (5-CSRTT) the shortest stimulus duration was used. For acoustic startle and prepulse inhibition (PPI) responses, data was pooled over all tested startle intensities, prepulse intensities and inter-stimulus intervals, to have an unbiased assessment since studies show conflicting results across the range of startle and prepulse intensities (Baker et al., 2010; Braat et al., 2015; Ding et al., 2014; Hodges et al., 2019; Michalon et al., 2012; Naviaux et al., 2015; Zhang et al., 2014). For olfactory sensitivity tasks the data from the lowest concentration that was still detectable by WT animals was used. In the olfactory habituation-dishabituation task all first presentations, excluding water, were pooled. In the gap-crossing task, the data from gap-distances between five and six cm were pooled. For task assessing novelty recognition (social or object novelty recognition) data was only extracted when a ratio or index was reported, as the time spent interacting with the novel object/animal is only informative relative to the time spent interacting with the familiar object/animal. For locomotor activity, whenever available, only the first 30 minutes of exploration were extracted.

Meta-analysis

The meta-analysis was performed using comprehensive meta-analysis (CMA, v.3.3, Biostat Inc., Englewood, NJ, USA). For most outcome measures Hedge's G standardized mean differences (SMD) were used as the effect size measure. For the outcome measure audiogenic seizures, odds ratios were calculated. Because of anticipated heterogeneity, the effect sizes were pooled using a random effects model. Overall SMD were reported with 95% confidence intervals (CI). I^2 was used to assess statistical heterogeneity (*i.e.*, variation across studies due to heterogeneity rather than chance (Higgins and Thompson, 2002).

To further explore heterogeneity, subgroup analyses were performed. Subgroup analysis was only performed when there were at least 10 comparisons, from 5 unique studies for each subgroup. The original protocol listed only 5 comparisons from 3 unique studies, but based on the advice from the SYRCLE institute we increased these numbers to increase the power of the subgroup analyses. The effects of sex were explored by comparing studies only using male animals, with mixed sexes studies, using either females or both males and females, since there was not enough data to analyse females and male-female combined data separately. The effect of age was explored by grouping the age of experimental animals into juvenile (<6 weeks), adolescent (mice: 6-9 weeks; rats: 6-21 weeks) and adult (mice: >9 weeks, rats: > 21 weeks) (Adriani et al., 2004; Ghasemi, Asghar; Sajad, 2021; Semple et al., 2013; Sengupta, 2013). When an age-range was reported, the study was grouped into the age category the majority of the range belonged to. In cases were only the age at the start of testing was reported, this same age was taken when consecutive tests were performed. Genetic background effects were tested for C57/Bl6(J&N), FVB and FVBx129. Additionally, the effects of single vs group housing, being littermates or not and the KO vs KO2 mouse model were tested. Subgroup analyses were not performed on the other characteristics that were extracted because of a lack of data; this includes species (rat vs mouse) which was pre-specified in the initial protocol as a subgroup analysis factor. The number of behavioural tasks in a study was also prespecified as a subgrouping factor. However, during the extraction of the characteristics it turned out to be a complex outcome to extract due to various reasons, including the difficulty of defining when different phases of one task become separate tasks (*e.g.*, initial learning and reversal learning in the Morris water maze), missing information of whether or not tests were performed in different batches of animals, and uncertainty about how these would affect the meaning of this outcome. Thus, this characteristic was not taken into the meta-analyses as a subgrouping factor.

To test for differences between subgroups we calculated the confidence interval of the difference between the subgroups. Whenever three subgroups were compared, Bonferroni corrections were applied to correct for multiple comparisons. P-values lower than 0.05 were considered statistically significant.

Sensitivity analysis

A sensitivity analysis was performed to check if methodological or experimental differences reported between the studies could be skewing the main effect and should be considered separately. For this, the main effects of the meta-analysis when including all the studies were compared to the main effects when taking out those studies that reported different methodologies (*e.g.*, open field for more than 30 minutes) or did not explicitly report details that were assumed at the extraction phase. If the main effect remained unchanged after the removal of those studies, it was implied that those methodological differences were not dragging the meta-analysis main effect, therefore they could remain included. Experimental differences included, for example, assuming the error bars represented SEM when not specified, tests or stimuli with different time lengths, whether data was pooled over stimuli or time, etc. See Supplementary file 7 for more details.

Publication bias assessment

Two different analyses were performed in parallel to assess whether meta-analyses showed significant asymmetry in the funnel plot and thus possibly suffered from publication bias, namely Egger's regression and Duval and Tweedie (Duval and Tweedie, 2000) trim and fill analysis (Stata Statistical Software, SE17, StataCorp LLC, College Station, TX). For both methods, the effect size estimate Hedges' G and sample-size based precision estimate $1/\sqrt{N}$ were used as it has been suggested that SE-based precision estimates cause distortion of SMD funnel plots (Wenstedt et al., 2021).

First, the Egger regression test was performed (Egger et al., 1997). This test is based on a simple linear regression and it can only identify small-study effects. In case of no publication bias, the regression line would cross the zero of the standard normal deviate (*i.e.*, precision estimate) in the y-axis.

Secondly, for the Duval and Tweedie, funnel plots were created where the effect sizes were plotted on the x-axis against 1/√N as a measure of precision on the y-axis (Zwetsloot et al., 2017). If there is no publication bias, studies are expected to spread equally across both sides of the overall effect size with larger deviations from the overall effect as the precision (*i.e.*, sample size) of the study decreases.

All data is publicly available in supplementary files via the OSF repository (https://osf.io/d2cbx/), this includes all the extracted data (Supplementary file 8), the statistical results of the meta-analysis (Supplementary file 9), the statistical results of the subgroup analyses (Supplementary file 10) and the statistical results of the publication bias analysis (Supplementary file 11). Additionally, on this repository the methods and results of the behavioural experiments we performed in our own lab can also be found.

Results

Search results

In total, 5065 records were retrieved through database screening. After duplicate removal 3414 unique records were scanned for eligibility. Via title and abstract screening, 374 records were selected for full-text assessment. Of those, 265 articles were found to be eligible for inclusion in the systematic review and risk of bias assessment, together with one study that was found by scanning the reference lists of included articles. From the 266 studies of the systematic review, 15 studies were excluded from the meta-analyses because they did not contain the right data (Fig. 1). Thus, 251 studies were included in the meta-analysis, as the minimum of five independent studies was reached for every behavioural domain. The digital object identifiers (DOIs) of all included studies can be found in the characteristics table (Supplementary file 3).

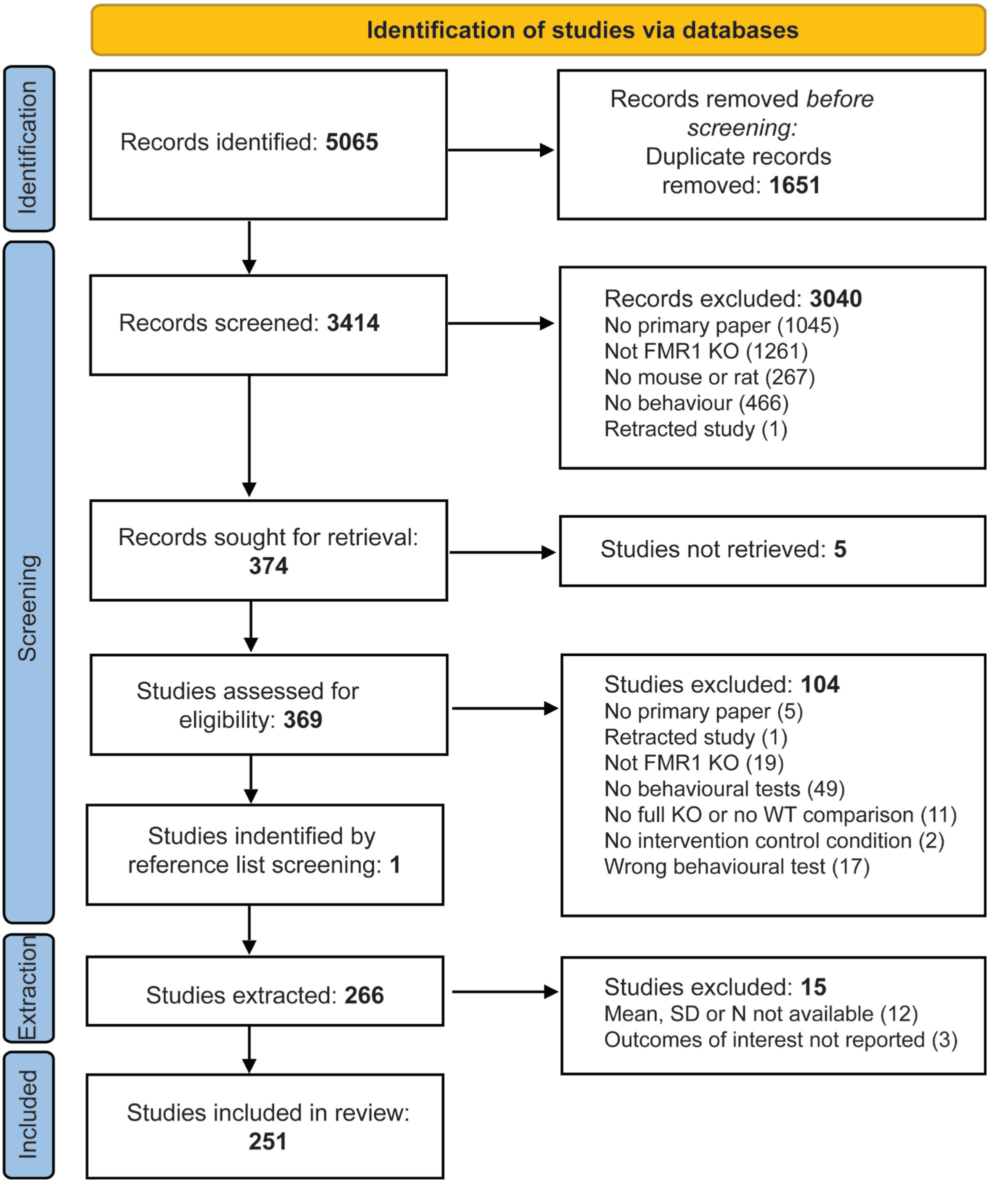

Identification of studies via databases
Identification
Screening
Extraction
Included
Records identified: 5065
Records removed *before screening:* Duplicate records removed: 1651
Records screened: 3414
Records excluded: 3040
No primary paper (1045)
Not FMR1 KO (1261)
No mouse or rat (267)
No behaviour (466)
Retracted study (1)
Records sought for retrieval: 374
Studies not retrieved: 5
Studies assessed for eligibility: 369
Studies excluded: 104
No primary paper (5)
Retracted study (1)
Not FMR1 KO (19)
No behavioural tests (49)
No full KO or no WT comparison (11)
No intervention control condition (2)
Wrong behavioural test (17)
Studies indentified by reference list screening: 1
Studies extracted: 266
Studies excluded: 15
Mean, SD or N not available (12)
Outcomes of interest not reported (3)
Studies included in review: 251

Fig 1. Study flowchart. All behavioural categories reached the minimum number of studies needed for meta-analysis, therefore all studies included in the systematic review were also included in the meta-analysis.

Study Characteristics

The characteristics of all included studies can be found in supplementary file 3. From the 266 included studies 252 used mice, 10 used rats and one study used both rats and mice. Of the studies performed in mice, the C57BL/6 background was the most frequently used background (151), but also FVB (80) and FVBx129 (9) were used frequently. Twenty-five studies used other backgrounds and six studies did not report the genetic background of the mice. In rat studies, both Sprague-Dawley (7) and Long Evans (4) backgrounds were used. Data was reported either specifically for males (215) and females (20), or for the two sexes combined (23). The sex of the animals was not specified in 22 studies. The majority of studies used adult animals (173), followed by juvenile (63) and adolescent (42) animals. In 23 studies, the age of the animals was not reported. From the studies using mice, 179 used the first-generation KO, 19 used the second-generation KO (KO2) and 65 did not specify which model was used. Most studies tested KO and WT animals as littermates (160), but in 33 studies control animals were not littermates and in 74 studies it was not reported. In most studies animals were group-housed (179), 26 of which used housing with mixed genotypes, 11 with the same genotype and for 142 studies it was unknown how the groups were composed. In 18 studies animals were individually housed during experiments and 123 studies did not report on housing conditions. The majority of studies did not specify whether behavioural tests were performed during the light phase or dark phase. From the studies that did report the testing phase, 116 performed tests during the light phase, 11 during the dark phase and three performed 24h recordings, thus including both light phases.

None of the experimental or methodological differences tested for in the sensitivity analysis affected the main effect in any of the meta-analyses.

Study Quality

A risk of bias assessment was performed according to the SYRCLE's RoB tool (Hooijmans et al., 2014) in a random subset of the included articles (Supplementary file 4). Overall, the risk of bias in these articles was unclear (Fig. 2). Blinded execution was reported in 49% of the articles and 58% of the studies assessed the outcomes blinded for genotype, while one study reported to not be blinded (2%). Except for one study stating that outcome assessment was not performed in a randomized order, none of the studies mentioned randomization of outcome assessment. Outcome data was incomplete in four studies (9%) and it was unclear whether data was complete in 78% of assessed studies. Five studies (11%) did not report on all the outcomes presented in the methods section.

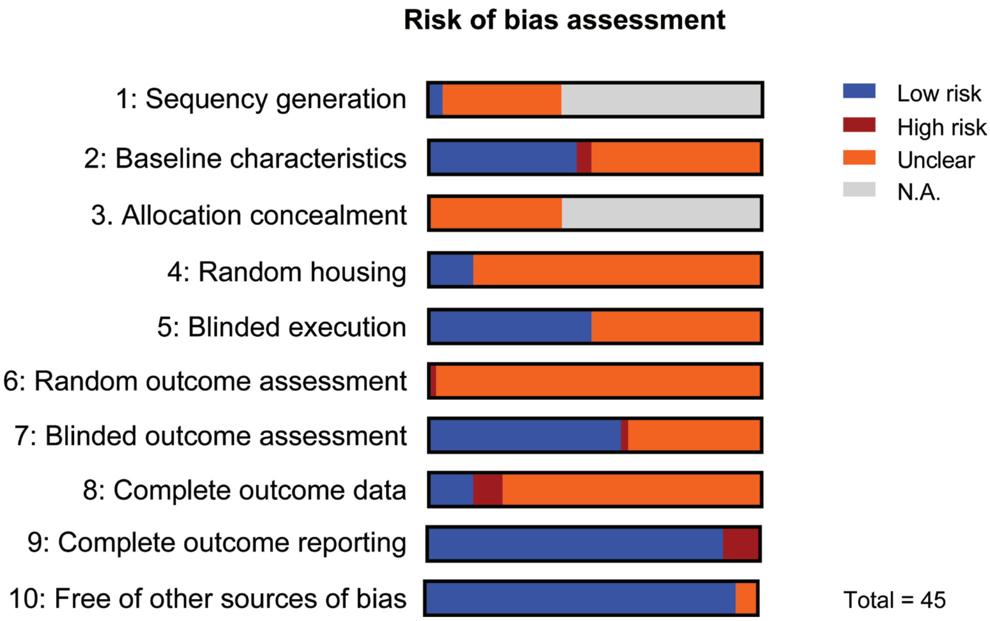


Fig 2. Risk of bias assessment outcomes. Risk of bias assessment was performed with SYRCLE's risk of bias assessment tool. Item 1 and 3 were only used in studies with an intervention, in which was assessed if animals were randomly assigned to the control condition. Risk of bias analysis was performed on a random sample of 45 studies by two independent assessors.

Locomotion

The meta-analysis comprised 176 comparisons out of 125 independent studies. A total of 2331 WT and 2299 KO animals were included in the analysis. The most frequently used behavioural test to assess locomotion was the open field test (145), followed by the three-chamber test (6), Novel Object Recognition Test (NORT) training phase (3), actimetry cages (3),

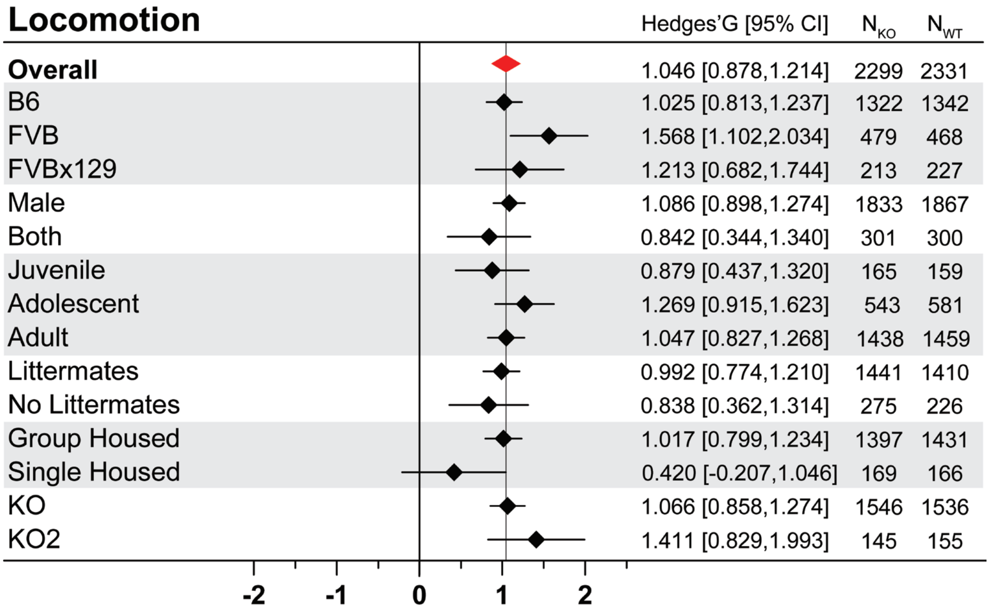


Fig 3. The effect of *Fmr1* KO on locomotor activity. Subgroup analyses were performed for genetic background, sex, age, littermates, housing condition and KO line. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N^{KO} and N^{WT}) included in each of the comparisons.

Morris water maze (2), home cage activity (1), active place avoidance (1), elevated plus maze (1), plus-shaped water maze (1) and visual cliff (1).

Ninety-two comparisons had a point estimate significantly larger than zero, four comparisons had a point estimate significantly smaller than zero and 80 comparisons did not significantly deviate from zero. Overall analysis showed that *Fmr1* KO animals have a significant increase in locomotor activity compared to WT controls (SMD 1.046 [0.878, 1.214], $P < 0.001$, $I^2 = 85.4$, Fig. 3, Supplementary file 5). The heterogeneity was considerable and remained unchanged after the subgroups analyses.

The genotype effect did not differ between genetic backgrounds (B6 vs FVB: t(132) = 2.08, *P* = 0.12; B6 vs FVBx129: t(121) = 0.64, *P* = 1.56; FVB vs FVBx129: t(51) = 0.99, *P* = 0.99), sexes (t(161) = 0.90, *P* = 0.37), age groups (Juvenile vs

Adolescent: t(54) = 1.35; *P* = 0.55); Juvenile vs Adult: t(121) = 0.67; *P* = 1.52; Adolescent vs Adult: t(145) = 1.04, *P* = 0.90), littermates and non-littermates (t(124) = 0.58, *P* = 0.56), the first and second generation KO (t(128) = 1.09, *P* = 0.28) nor single and group housed animals (t(114) = 1.77, *P* = 0.080).

Cognition

Conditioned Learning

The meta-analysis comprised 134 comparisons out of 83 independent studies. A total of 1752 WT and 1791 KO animals were included in the analysis. The most frequently used behavioural test to assess conditioned learning was fear conditioning (70), followed by passive avoidance (35), discrimination learning (13), active avoidance (6), operant conditioning (6), conditioned place preference (3) and conditioned taste aversion (1).

Sixty-four of the comparisons had a point estimate significantly smaller than zero, three comparisons had a point estimate significantly larger than zero and 67 comparisons did not significantly deviate from zero. Overall, *Fmr1* KO animals show a significant decrease in conditioned learning compared to WT controls (SMD -0.862 [-1.023, -0.702], *P* < 0.001, I^2 = 80.3 Fig. 4, Supplementary file 5), however there was high heterogeneity which did not decrease with subgroup analysis.

The difference between KO and WT animals seemed larger in the FVBx129 and FVB backgrounds compared to the B6 background, although not significantly (B6 vs FVBx129: t(84) = 1.92, *P* = 0.18; B6 vs FVB: t(99) = 2.07, *P* = 0.12; FVB vs FVBx129: t(37) = 0.03, *P* = 2.91). The genotype effect was larger in animals that were not littermates, compared to animals that were littermates (t(93) = 2.56, *P* = 0.012). The genotype effect did not differ between sexes (t(121) = 0.98, *P* = 0.33), nor age groups (Juvenile vs Adolescent: t(28) = 0.28, *P* = 2.34; Juvenile vs Adult: t(107) = 1.01, *P* = 0.94; Adolescent vs Adult: t(109) = 1.24, *P* = 0.65).

Spatial Cognition

The meta-analysis comprised 69 comparisons out of 42 independent studies. A total of 725 WT and 752 KO animals were included in the analysis. The most frequently used behavioural test to assess spatial memory was object location memory (24), followed by the Morris water maze (17), categorical spatial processing task (6), plus-shaped water maze (5), radial maze (5), non-match to place learning (4), y-maze (3), Barnes maze (2), E-maze (1), Hebb-William maze (1) and metric change in the NORT (1).

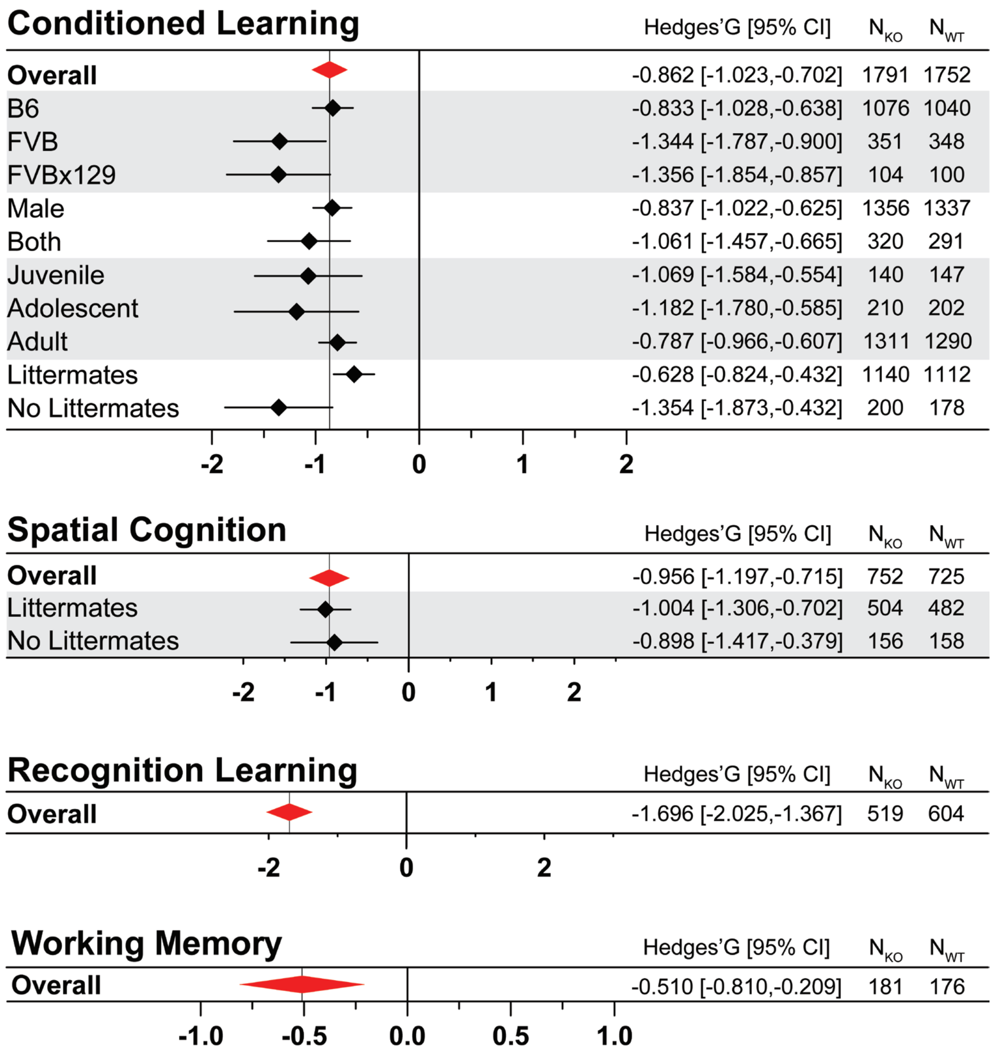
Thirty-three out of these comparisons had a point estimate significantly smaller than zero, two had a point estimate significantly larger than zero and 34 comparisons did not deviate from zero. *Fmr1* KO animals show a robust and significant impairment in spatial cognition compared to WT controls (SMD -0.956 [-1.197, -0.715], $P < 0.001$, $I^2 = 79.1$, Fig. 4, Supplementary file 5). Heterogeneity did not reduce with subgroup analysis.

The genotype effect did not differ between animals that were littermates or no littermates ($t(58) = 0.35$, $P = 0.73$).

Recognition Learning

The meta-analysis comprised 53 comparisons out of 34 independent studies. A total number of 604 WT and 519 KO animals were included in the analysis. All studies used the NORT, two of which used the temporal order version of the NORT.

Forty out of these comparisons had a point estimate significantly smaller than zero and 13 comparisons had a point estimate not significantly different from zero. *Fmr1* KO animals show a robust and significant impairment in recognition memory compared to WT controls (SMD -1.696 [-2.025, -1.367], $P < 0.001$, $I^2 = 82.5$, Fig. 4, Supplementary file 5). Heterogeneity did not reduce with subgroup analysis.



Conditioned Learning	Hedges'G [95% CI]	N_{KO}	N_{WT}
Overall	-0.862 [-1.023,-0.702]	1791	1752
B6	-0.833 [-1.028,-0.638]	1076	1040
FVB	-1.344 [-1.787,-0.900]	351	348
FVBx129	-1.356 [-1.854,-0.857]	104	100
Male	-0.837 [-1.022,-0.625]	1356	1337
Both	-1.061 [-1.457,-0.665]	320	291
Juvenile	-1.069 [-1.584,-0.554]	140	147
Adolescent	-1.182 [-1.780,-0.585]	210	202
Adult	-0.787 [-0.966,-0.607]	1311	1290
Littermates	-0.628 [-0.824,-0.432]	1140	1112
No Littermates	-1.354 [-1.873,-0.432]	200	178

Spatial Cognition	Hedges'G [95% CI]	N_{KO}	N_{WT}
Overall	-0.956 [-1.197,-0.715]	752	725
Littermates	-1.004 [-1.306,-0.702]	504	482
No Littermates	-0.898 [-1.417,-0.379]	156	158

Recognition Learning	Hedges'G [95% CI]	N_{KO}	N_{WT}
Overall	-1.696 [-2.025,-1.367]	519	604

Working Memory	Hedges'G [95% CI]	N_{KO}	N_{WT}
Overall	-0.510 [-0.810,-0.209]	181	176

Fig 4. The effect of *Fmr1* KO on cognition. Meta-analyses were performed in the category of conditioned learning, spatial cognition, recognition learning and working memory. Subgroup analyses were performed for genetic background, sex, age, and/or littermates. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N^{KO} and N^{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

Working Memory

The meta-analysis comprised 15 comparisons out of 13 independent studies. A total of 183 WT and 187 KO animals were included in the analysis. The most frequently used behavioural task to assess working memory was spontaneous alternations in the Y-maze (6) and the T-maze (6), followed by working memory errors in radial arm maze learning (2), delayed non-match to place learning (1) and serial reversals in the Morris water maze (1).

Three out of these comparisons had a point estimate significantly smaller than zero, while 12 comparisons did not deviate from zero. *Fmr1* KO animals show a significant impairment in working memory compared to WT controls (SMD -0.510 [-0.810, -0.209], *P* = 0.001, I^2 = 49.7% Fig. 4, Supplementary file 5).

Repetitive behaviour

Low order repetitive behaviour

The meta-analysis comprised 87 comparisons out of 53 independent studies. A total of 1063 WT and 1098 KO animals were included in the analysis. The most frequently used behavioural test to assess low order repetitive behaviour was the marble burying test (42), followed by spontaneous behaviour in the open field (33), fear conditioning (3), three-chamber (2), y-maze (1) or elevated plus maze (1), block chew test (2) and the nose-poke assay (2).

Thirty-six out of these comparisons had a point estimate significantly larger than zero, nine comparisons had a point estimate significantly smaller than zero and 42 comparisons did not deviate from zero. *Fmr1* KO animals show a significant increase in low order repetitive behaviours compared to WT controls (SMD 0.572 [0.356, 0.789], *P* < 0.001, I^2 = 82.4, Fig. 5, Supplementary file 5), however there was considerable heterogeneity which did not decrease with subgroup analysis.

The genotype effect did not differ between genetic backgrounds (B6 vs FVB: t(69) = 0.25, *P* = 0.81), sex (t(82) = 0.78, *P* = 0.44), age groups (Juvenile vs Adolescent: t(23) = 0.51, *P* = 1.84, Juvenile vs Adult: t(70) = 0.81, *P* = 1.27; Adolescent vs Adult: t(73) = 1.56, *P* = 0.34), nor sexes (t(82) = 0.78, *P* = 0.44).

Cognitive Flexibility

The meta-analysis comprised 30 comparisons out of 23 independent studies. A total of 352 WT and 361 KO animals were included in the analysis. The most frequently used behavioural test to assess cognitive flexibility was reversal in the Morris water maze (10), followed by discrimination learning reversal (4), plus-shaped water maze reversal (3), y-maze reversal (3), passive avoidance extinction (3), active avoidance extinction (3), operant conditioning extinction (1), fear conditioning extinction (1), E-maze reversal (1) and 5-CSRRT reversal (1).

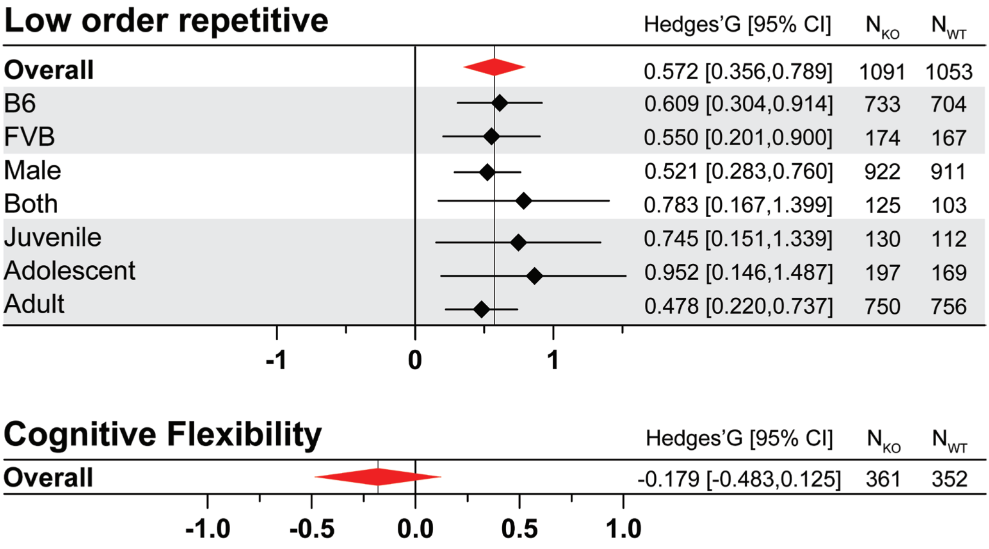


Fig 5. The effect of *Fmr1* KO on repetitive and restricted behaviour. Meta-analyses were performed in the category of low order repetitive behaviour and cognitive flexibility. Subgroup analyses were performed for genetic background, sex and age. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N^{KO} and N^{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

Eight out of these comparisons had a point estimate significantly smaller than zero, four had a point estimate significantly larger than zero and 18 comparisons did not deviate from zero. *Fmr1* KO animals do not show a cognitive flexibility deficit (SMD -0.179 [-0.483, 0.125], *P* < 0.249, I^2 = 75.0, Fig. 5, Supplementary file 5).

Social behaviour

Sociability

The meta-analysis comprised 107 comparisons out of 69 independent studies. A total of 1424 KO and 1399 WT animals were included in the analysis. The most frequently used behavioural task to assess sociability was the three-chamber test (67), followed by the direct social interaction test (23), partition test (11), tube co-occupancy test (2), resident-intruder test (2), Eco-HAB (1) and the social conditioned place preference test (1).

Thirty-six out of these comparisons had a point estimate significantly smaller than zero, 10 comparisons had a point estimate significantly larger than zero and 61 studies did not deviate from zero. *Fmr1* KO animals show a significant decrease in sociability compared to WT controls (SMD -0.368 [-0.546, -0.189], *P* < 0.001, I^2 = 81.1, Fig. 6, Supplementary file 5), however there was a high degree of heterogeneity which did not reduce in subgroup analysis.

The genotype effect was significantly larger in studies using only male animals compared to studies using both sexes, in which the effect was also not significantly different from zero (SMD 0.143 [-0.315 0.601], t(99) = 2.19, *P* = 0.030). The genotype effect did not differ between genetic backgrounds (B6 vs FVB: t(86) = 0.62, *P* = 0.53), age groups (Juvenile vs Adolescent: t(25) = 0.8673, *P* = 1.54; Juvenile vs Adult: t(83) = 0.72, *P* = 1.42; Adolescent vs Adult: t(84) = 0.10, *P* = 2.76), littermates and non-littermates (t(78) = 1.39, *P* = 0.17).

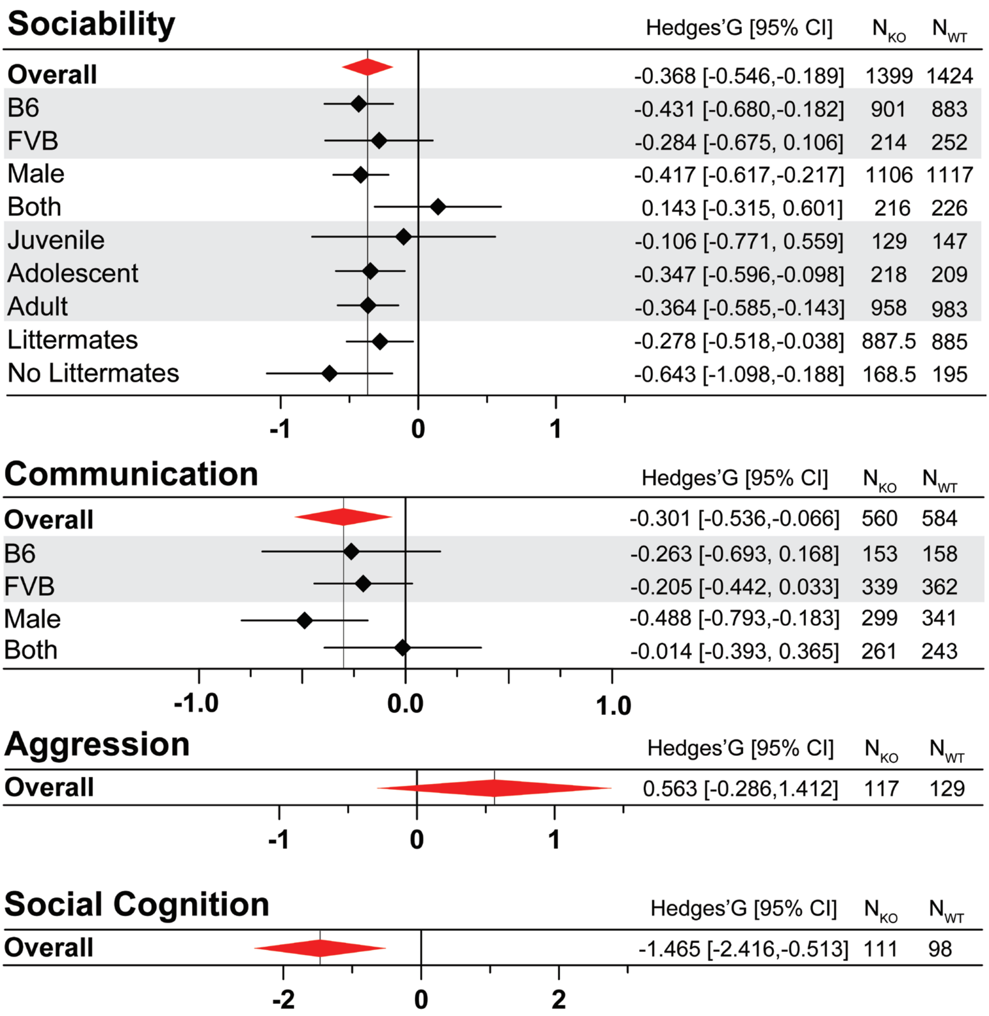


Fig 6. The effect of *Fmr1* KO on social behaviour. Meta-analyses were performed in the category of sociability, communication, aggression and social cognition. Subgroup analyses were performed for genetic background, sex, age and littermates. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N^{KO} and N^{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

Communication

The meta-analysis comprised 35 comparisons out of 21 independent studies. A total of 584 WT and 560 KO animals were included in the analysis.

Most USVs were isolation-induced (22) or socially-induced (10). USVs were also recorded in the resident-intruder test (2) and the open field (1).

Ten out of these comparisons had a point estimate significantly smaller than zero, four comparisons had a point estimate larger than zero and 21 comparisons did not deviate from zero. *Fmr1* KO animals show a significant communication deficit (SMD -0.301 [-0.536, -0.066], *P* = 0.127, I^2 = 72.1, Fig. 6, Supplementary file 5). The overall heterogeneity did not reduce in subgroup analysis.

The genotype effect was only present in studies using only males (-0.488 [-0.793, -0.183]), and not in studies using both sexes (-0.014 [-0.393, 0.365]), although the difference between the sexes was not significant (t(33) = 1.91, *P* = 0.065). The genotype effect did not differ between genetic backgrounds (t(27) = 0.23, *P* = 0.82).

Aggression

The meta-analysis comprised 10 comparisons out of six independent studies. A total of 129 WT and 117 KO animals were included in the analysis. The most frequently used test to assess aggressive behaviour was the direct social interaction task (6) followed by the tube test (3) and the dominance hierarchies (1).

Five out of these comparisons had a point estimate significantly larger than zero, two had a point estimate significantly larger than zero and three comparisons did not significantly deviate from zero. *Fmr1* KO animals did not show enhanced aggression (SMD 0.563 [-0.286, 1.412], *P* = 0.194, I^2 = 89.6, Fig. 6, Supplementary file 5).

Social Cognition

The meta-analysis comprised 10 comparisons out of seven independent studies. A total of 98 WT and 111 KO animals were included in the analysis. All assessments of social cognition were performed in the three-chamber test.

Eight out of these comparisons had a point estimate significantly smaller than zero and two comparisons did not deviate from zero. *Fmr1* KO animals showed a consistent significant reduction in social cognition (SMD -1.465 [-2.416, -0.513], *P* = 0.003, I^2 = 87.9, Fig. 6, Supplementary file 5).

Anxiety

The meta-analysis comprised 136 comparisons out of 96 independent studies. A total of 1838 WT and 1882 KO animals were included in the analysis. The most frequently used behavioural test to assess anxiety was the open field (58), followed by the elevated plus maze (36), the light-dark test (32), the elevated zero maze (6), the successive alleys maze (2), the mirrored chamber (1) and the platform test (1).

Fifty-two out of these comparisons had a point estimate significantly smaller than zero, seven comparisons had a point estimate significantly larger than zero and 77 studies did not deviate from zero. *Fmr1* KO animals show a significant decrease in anxiety compared to WT controls (SMD -0.555 [-0.692, -0.419], *P* < 0.001, I^2 = 75.1, Fig. 7, Supplementary file 5), The overall heterogeneity did not decrease with subgroup analysis.

The genotype difference was smaller in juvenile compared to adolescent and adult animals, although this difference was not significant (Juvenile vs Adolescent: t(39) = 1.29, *P* = 0.61; Juvenile vs Adult: t(96) = 1.26, *P* = 0.63; Adolescent vs Adult: t(115) = 0.15, *P* = 2.65). Similarly, although the effect was larger in studies using both sexes, this difference did not reach statistical significance (t(127) = 1.73, *p* = 0.086). The difference between KO and WT animals was not affected by genetic background (B6 vs FVB: t(98) = 0.97, *P* = 1.01; B6 vs FVBx129: t(92) = 1.12, *P* = 0.71; FVB vs FVBx129: t(40) = 1.66, *P* = 0.32), nor littermates and non-littermates (t(106) = 0.98, *P* = 0.33).

Attention

The meta-analysis comprised seven comparisons out of five independent studies. A total of 84 WT and 91 KO animals were included in the analysis.

All assessments of attention were performed in the 5-choice serial reaction time task. One of the comparisons had a point estimate significantly smaller than zero, one had a point estimate significantly larger than zero and five comparisons did not deviate significantly from zero. *Fmr1* KO animals did not show an attention deficit (SMD 0.064 [-0.555, 0.683], *P* = 0.839, I^2 = 75.5, Fig. 7, Supplementary file 5).

Startle and prepulse inhibition

Acoustic Startle

The meta-analysis comprised 56 comparisons out of 40 independent studies. A total of 883 WT and 866 KO animals were included in the analysis. Nineteen out of these comparisons had a point estimate significantly smaller

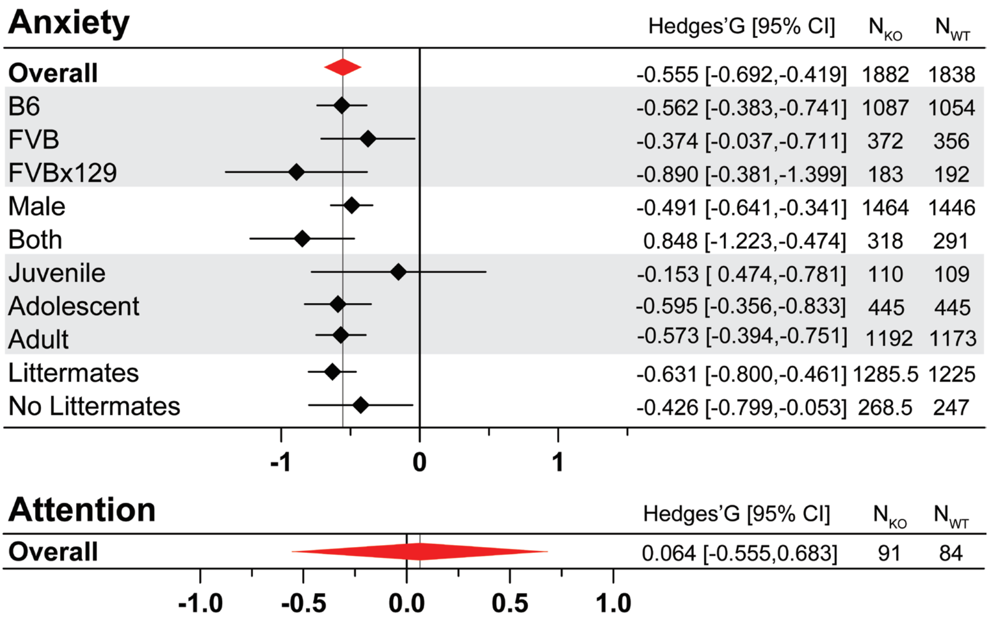


Fig 7. The effect of *Fmr1* KO on anxiety and attention. Subgroup analyses were performed for genetic background, sex, age and littermates. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N^{KO} and N^{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization.

than zero, seven comparisons had a point estimate significantly larger than zero and 30 comparisons did not deviate from zero. *Fmr1* KO animals show a significantly decreased acoustic startle compared to WT controls (SMD -0.335 [-0.591, -0.079], *P* = 0.010, I^2 = 84.9, Fig. 8, Supplementary file 5).

The startle deficit was present in mice with a FVB background (-0.838 [-1.242, -0.434]), but not in mice with a B6 background (-0.045 [-0.471 0.381], t(36) = 2.65, *P* = 0.012). The overall heterogeneity did not reduce with subgroup analysis.

Prepulse Inhibition

The meta-analysis comprised 46 comparisons out of 30 independent studies. A total of 613 WT and 598 KO animals were included in the analysis. Twenty out of these comparisons had a point estimate significantly larger than zero and 26 comparisons did not deviate from zero. *Fmr1* KO animals show a significantly increased prepulse inhibition compared to WT controls (SMD 0.601 [0.403, 0.799], *P* < 0.001, I^2 = 65.2, Fig. 8, Supplementary file 5). Although effects in the opposite direction were not found, the heterogeneity was still considerable. The genotype effect did not differ between genetic backgrounds (t(29) = 0.15, *P* = 0.89).

Sensory sensitivity

The meta-analysis comprised 41 comparisons out of 26 independent studies. A total of 525 WT and 484 KO animals were included in the analysis. The most frequently used behavioural test to assess sensory sensitivity was the hot plate (16), followed by chemically-induced pain (5), odour habituation-dishabituation test (4), odour discrimination (3), buried food test (2), von Frey test (2), gap crossing task (2), olfactory sensitivity test (2), whisker-dependent texture discrimination (2), visual cliff test (1), texture NORT (1) and shock sensitivity (1). Eight out of these comparisons had a point estimate significantly smaller than zero, one comparison had a point estimate significantly larger than zero and 32 comparisons did not deviate from zero. *Fmr1* KO animals show a significantly decreased sensory sensitivity compared to WT controls (SMD -0.412 [-0.586, -0.239], $p < 0.001$, $I^2 = 46.7$, Fig. 8, Supplementary file 5). The overall heterogeneity did not decrease with subgroup analysis.

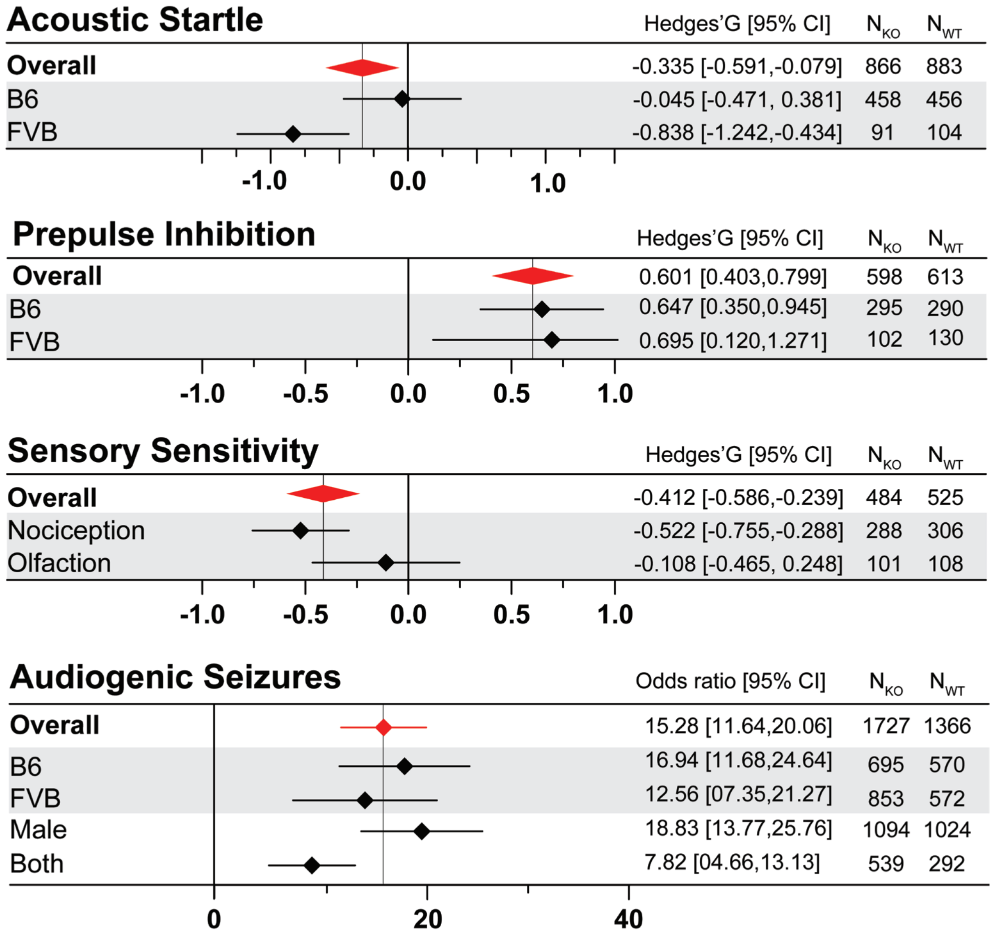


Fig 8. The effect of *Fmr1* KO on sensory processing. Meta-analyses were performed in the category of acoustic startle, prepulse inhibition, sensory sensitivity and audiogenic seizures. Subgroup analyses were performed for genetic background, sex and sensory modality. Data are presented as Hedges' G standardized mean difference and 95% confidence interval. The last two columns report the number of animals per genotype (N^{KO} and N^{WT}) included in each of the comparisons. X-axis limits differ per meta-analysis to optimize subgroup error band visualization. Note that the results of the audiogenic seizures (bottom panel) are expressed in odds ratio.

The sensory sensitivity deficit seemed to be stronger for nociception compared to olfaction, though not significantly (t(31) = 1.96, *p* = 0.059).

Audiogenic seizures

The meta-analysis comprised 98 comparisons out of 40 independent studies. A total of 1376 WT and 1737 KO animals were included in the analysis. Sixty-three out of these comparisons had a point estimate significantly larger than zero, 35 comparisons did not deviate from zero. *Fmr1* KO animals show a robust and significant increased sensitivity for audiogenic seizures (Odds ratio 15.280 [11.643, 20.055], *P* < 0.001, I^2 = 11.7, Fig. 8, Supplementary file 5). The overall heterogeneity did not decrease with subgroup analysis.

The genotype effect was larger in studies using only male animals compared to studies using both sexes (t(92) = 2.47, *P* = 0.015). The genetic background did not affect seizure sensitivity (B6 vs FVB: t(83) = 0.73, *P* = 0.46.

Publication bias

Publication bias was assessed through funnel plot's asymmetry according to Egger's regression test for small-study effects supplemented with Duval and Tweedie trim and fill analysis. The meta-analysis for audiogenic seizures was performed using odds ratio and was therefore assessed for publication bias only with a trim and fill analysis.

Inspection of the funnel plots did not reveal asymmetry with either the Egger test nor Duval and Tweedie test for the anxiety, aggression, conditioned learning, cognitive flexibility, communication, locomotion, PPI, sensory sensitivity, sociability, and social cognition (Supplementary file 11).

Egger's regression test indicated bias for three behavioural categories: acoustic startle (P = 0.005), recognition learning (P = 0.012), and spatial cognition (P = 0.040) (Supplementary file 11).

The Duval and Tweedie test showed funnel plot asymmetry for two categories, namely acoustic startle and low order repetitive behaviour. For the acoustic startle response, studies showing increased startle response by the KO animals compared to WT were underrepresented. This resulted in 15 imputed studies and an adjusted effect size estimate of 0.045 [-0.249, 0.339] (Supplementary file 6). In the low order repetitive behaviour, studies showing decreased repetitive results were underrepresented leading to imputing 27 extra studies resulting in an adjusted effect size to -0.031 [-0.370, 0.307] (Supplementary file 6). For these two categories, the direction of the effect size changed after adjusting for the trim and fill analysis. These results should therefore be cautiously interpreted as marginal effects could be inflated by publication bias.

Additionally, the trim and fill analysis using odds ratios for the audiogenic seizures also showed funnel plot asymmetry. Twenty-four extra studies were added and gave an adjusted effect size of 3.917 [3.213, 4.774] (Supplementary file 6). The effect size direction remained §the same after adjustment.

The discrepancies shown by these two publication bias analysis methods could be explained by the different methodologies they use. However, both methods indicated a significant overestimation of the genotypic effect in the acoustic startle response due to publication bias.

Discussion

In the current systematic review and meta-analysis, we aimed to shed light on the behavioural profile of the *Fmr1* KO model, how it matches the clinical manifestation of FXS and which experimental factors might explain the heterogeneity of results seen in literature. We were able to include a large body of literature, which allowed us to perform meta-analyses in all relevant behavioural categories; however, in preclinical meta-analyses there is a trade-off between power and heterogeneity, which makes correct interpretation of the overall effect more complex. Irrespective of the overall effects found, this meta-analysis underscores the large inconsistencies between studies with effects being replicated in less than 50% of the independent comparisons in 10 out of 14 categories. This heterogeneity could represent true between-study variation in study design and experimental conditions (*i.e.*, phenotypic flexibility due to environmental diversity), but this was hard to assess due to the poor reporting of experimental factors. Additionally, low sample sizes and suboptimal research practices are likely to contribute to the low replicability of the phenotypes as studies with higher effect sizes showed more consistent results. Both incomplete reporting of experimental methods and conditions, and underpowered studies are common problems in behavioural preclinical neuroscience research (Sena et al., 2014). This meta-analysis stresses the need for improvements, not only regarding the *Fmr1* KO, but for animal research in the preclinical field as a whole.

With the estimated overall effect sizes that resulted from the meta-analysis, we were able to look at achieved power and required sample sizes in the various behavioural categories. Based on the estimated overall effect size, required sample sizes for sociability and anxiety would be more than 100 animals per genotype in order to reach a statistical power of 0.8. Additionally, when calculating achieved power based on the estimated overall effect size and the average sample size, only recognition learning and social cognition reach a power of at least 0.8. However, computed achieved power does not match perfectly with the percentage of studies replicating certain effects, indicating that power is probably not the only factor causing the inconsistency of results. Also, these post-hoc calculations must be interpreted carefully, since due to the diverse experimental designs and research practices across studies, they cannot be translated to specific experimental settings. Nevertheless, insufficient power and sample sizes should be addressed as contributing factors to the low replicability of results.

It has also been reported that the rigorous standardization of animal experiments can lead to behavioural findings that can be replicated only under the exact same environmental and experimental conditions, which limits the interpretation and replicability of results (Richter, 2017; Voelkl and Würbel, 2016; Wurbel, 2000). FXS, like most neuropsychiatric disorders, is a complex disease where patients show high variability of phenotypes in terms of their symptoms and their severity (Ciaccio et al., 2017; Jacquemont et al., 2014). Likewise, animal models have shown phenotypic flexibility and so, the inconsistency of results between preclinical studies may be partly explained by the restricted generalizability and accuracy of results from study to study. Incorporating controlled biological variation into animal experiments could increase the external validity of findings (Voelkl et al., 2020). In addition, multicentre studies (Inthout et al., 2016) or multi-batch studies (Karp et al., 2020) are recommended in order to increase the robustness of studies assessing behavioural phenotype of animal models as these experimental designs have proven to render more representative study samples which allows more generalizable results. This could contribute to higher consistency across findings and thus more conclusive results.

Despite the large heterogeneity, we found significant overall effects matching the direction of the clinical profile in the majority of behavioural categories (Table 1). However, no effects were found on cognitive flexibility, attention and aggression although patients show flexibility and attention deficits, and enhanced aggression (Table 1, in bold). Nevertheless, these meta-analyses which did not show effects had a relatively low number of studies and total number of animals, and sometimes large confidence intervals, so results should be interpreted carefully. On the other hand, the reduced anxiety and acoustic startle, and enhanced PPI found in the KO animals are even opposite to the symptoms seen in patients. Strikingly, in patients the prevalence of problems with attention (74-84%), aggression (90%) and anxiety (58-86%) are higher than the prevalence of ASD (30-50%) and epilepsy (10-20%) of which the social, repetitive and seizure phenotypes were captured in the KO animals (Ciaccio et al., 2017).

Table 1. Comparison of the meta-analysis findings to the clinical phenotype.

Behavioural category	**Meta-Analysis**	**Clinical Phenotype**
Locomotion	↑	↑[1]
Conditioned learning	↓	↓[1,2]
Spatial cognition	↓	↓[1,3]
Recognition learning	↓	↓[1,4]
Working memory	↓	↓[5]
Low order repetitive	↑	↑[6]
Cognitive flexibility	=	↓[5]
Sociability	↓	↓[6]
Communication	↓	↓[6]
Aggression	=	↑[1]
Social cognition	↓	
Anxiety	↓	↑[1]
Attention	=	↓[1,5]
Acoustic startle	↓	↑[7-10]
PPI	↑	↓[7-10]
Sensory sensitivity	↓	↓↑[11]
Audiogenic seizures	↑	↑[1]

Categories in which the findings of the meta-analysis do not match the clinical phenotype are printed in bold text. [1](Ciaccio et al., 2017), [2](Reeb-Sutherland and Fox, 2015), [3](MacLeod et al., 2010), [4](Kogan et al., 2009), [5](Schmitt et al., 2019), [6](Niu et al., 2017), [7](Berry-Kravis et al., 2009), [8](Frankland et al., 2004), [9](Hessl et al., 2009), [10](Yuhas et al., 2011), [11](Baranek et al., 2009).

There are multiple possible explanations for the phenotype mismatch between the meta-analysis and the clinical population considering anxiety, startle and PPI. True species-specific differences in the mechanisms and thus the way the disorder presents itself in rodents and humans may exist. Discrepancies in anxiety findings might also result from the challenging assessment and interpretation of this complex behaviour in rodents. For example, some drugs known to be anxiolytic in humans are ineffective or even anxiogenic in the open field test, the most frequently used anxiety test in this meta-analysis (Prut and Belzung, 2003), questioning its suitability to capture anxiety behaviour. Moreover, most animal experimental designs tend to measure novelty-induced anxiety instead of long-term anxiety, which would be closer to the clinical setting. It has been suggested that the discrepancy can also be explained by a dissociation of social and generalized anxiety (Liu and Smith, 2009). Indeed, social anxiety is well documented in human literature; however, in preclinical studies it is confounded with other behavioural outcomes (*e.g.*, sociability) therefore, it was not possible to assess the fitness of the *Fmr1* KO model for this specific construct. However, while social phobia is the most common form of anxiety in FXS patients (Cordeiro et al., 2011), 50% of the patients show also generalized anxiety and 40% of the patients show agoraphobia, for which the open field test could be considered a very suitable test. Dissociation of generalized and social phobia can therefore only partly explain the discrepancies in anxiety phenotypes. Contrary to anxiety, the assessment of acoustic startle and PPI has a greater level of similarity between species; however, the relevance of the auditory stimuli might differ between the species as they primarily rely on different senses. Compensatory upregulation of FMRP-associated proteins in the KO mice may underlie the opposite phenotypes (Frankland et al., 2004; Paylor et al., 2008), as double mutant mice lacking both *Fmr1* and FXR2 (*FMR1* autosomal homolog 2) show decreased levels of PPI (Spencer et al., 2006). Furthermore, for all phenotypes which do not match the clinical profile it is important to keep in mind that these differences could be the result of a mismatch in disease induction in the KO models and patients. In contrast to the human condition, in neither of these two KO models the loss of protein is induced via an increase in CGG repeats. As the hypermethylation and thus silencing of protein expression in patients was shown to happen only at approximately the 12th day of gestation (Willemsen et al., 2002), differences in protein expression during early development could cause potential differences between the models and the clinical population. *Fmr1* knock-in (KI) models with increased CCG repeat expansions have been developed (Bontekoe et al., 2001; Entezam et al., 2007), but are currently only used to study the premutation (55-200 repeats) associated with Fragile X Tremor and Ataxia Syndrome (FXTAS). Although the mice also show repeat instability and permutation expansions that can develop into full mutation expansion numbers (>200 repeats; Entezam et al., 2007), for unknown reasons these expansion numbers are not resulting in protein silencing in mice (Entezam et al., 2007; Zhao et al., 2019). Therefore, the KO models are currently the best option to study FXS. However, in view of construct validity future studies should also consider to unravel why full mutation expansion numbers do not lead to protein silencing in mice in order to overcome the hurdles in developing functional KI models with increased CCG repeat expansions.

To be able to use anxiety, startle response and PPI in therapeutic interventions, it is important to further understand the phenotype discrepancies to allow for better interpretation and translation of rodent findings to clinical predictions.

In addition to assessing overall genotype effects, an important goal of this meta-analysis was to gain insight into factors that could explain the heterogeneity of the results in literature. Most of the overall genotypic effects scored a heterogeneity >70%, indicating high variability of the genotype effect between studies. This was also suggested by the substantial percentage of studies that reported a different direction of the effect than the overall effect.

Overall, few significant subgroup effects were found which only changed effect sizes but not the direction of effects. Additionally, the heterogeneity of the meta-analyses as assessed by the I^2-value, did not decrease after performing the subgroup analyses. This includes the sex of the animal and the maternal genotype, which were expected to explain some of the variation based on the fact that FXS is an X-linked syndrome and earlier research showing differences between WT animals from WT or heterozygous dams

(Zupan et al., 2016; Zupan and Toth, 2008). Together, these findings suggest that overlooked experimental factors introduced variability to our results. We speculate that the light phase in which the animals were tested and whether animals were single or group-housed could be relevant given their biological significance. These factors were included in the characteristics' extraction, but were reported too infrequently to be able to test their effects. Reporting these details information is important, as for example enriched environments have shown to reverse some of the phenotypes (Li et al., 2020; Restivo et al., 2005). In addition, we would like to highlight the infeasibility of making a cross-species assessment given the low number of studies performed with rats. Furthermore, it is possible that the variety of behavioural tests used explains part of the heterogeneity. Tests could differ in sensitivity to pick up certain phenotypes, or they may assess different aspects of the same phenotype. Although an exploratory analysis for this hypothesis did not show any indication of differences between the various tests used in the category of anxiety, our dataset allows for this assessment also in the other behavioural categories. These future analyses could not only give insight into whether different tests might pick up subtly different phenotypes, but also whether the between-study heterogeneity differs between the various behavioural tests available.

Possibly, few effects were found as the assessed experimental factors do not affect the genotype effect independently, but interact among each other. Although the current analysis did not allow for assessing these interactions, current developments in complex modelling and machine learning would allow for extracting more information from the same data.

All in all, these results urgently call the preclinical research community to improve research practices and reporting to boost the quality of data to generate more meaningful and conclusive results; which also applies to systematic reviews and meta-analyses of preclinical studies (Hunniford et al., 2021). Since environmental factors are such a big driver of phenotypic variability, better reporting of experimental conditions is necessary to increase understanding of the true heterogeneity in results.

While systematic reviews and meta-analyses give comprehensive summaries of the existing literature, it is important to realize that they are not completely bias-free, as results of this meta-analysis are inherently dependent on the methodological decisions made to align the diverse datasets. For example, when due to phenotypic flexibility acoustic startle phenotypes may present themselves in different startle intensities across various experimental conditions, averaging within each study over all tested startle intensities could lead to an underestimation of effect sizes. Unfortunately, as only a single outcome can be extracted per study, these kinds of decisions are unavoidable, nevertheless all decisions made for the current study were carefully discussed to minimize any kind of bias that could mislead the interpretation of results. However, some methodological decisions, in particular the boundaries of the age categories, are rather arbitrary since there is no consensus on these thresholds in literature. There are also studies stating that mice reach adulthood only after three months of age (Flurkey et al., 2007), and age subgroup analysis using this threshold actually showed a significant age effect on the anxiety phenotype (data not shown). Because of the non-consensus about these thresholds in the field, reporting the actual ages should always be preferred over only reporting the developmental stage of the experimental animals. An additional limitation of meta-analyses is that there are currently no automated methods to perform the data and characteristics extraction, therefore they are prone to human error. However, when running a random sub-sample check we only found 3.5% of errors in the characteristics extraction; given the large size of the meta-analyses, these errors minimally altered the effect sizes and did not change any of the outcomes of subgroup analysis (*i.e.*, conclusions stayed the same).

Additionally, the quality of this meta-analysis is dependent on the quality of the data included. There are numerous accounts, including the quality assessment in this meta-analysis, highlighting the often poor reporting and flawed experimental design of many preclinical studies (Kilkenny et al., 2009; Landis et al., 2012). This includes the prevalent lack of use, or reporting, of blinding and randomization, and poor research practices such as inadequate statistical tests and p-HARking (formulating the Hypothesis After gathering Results) (Bishop, 2019). The risk of bias assessment of the current systematic review and meta-analysis showed suboptimal reporting. Only 48.9% of the screened studies reported being 'blinded' for the experimental groups when conducting experiments while only 59.5% reported it for assessing the outcome. In 86.7% of studies, the housing arrangement of the animal subjects was unclear (*e.g.*, group housed with mixed genotypes). Strikingly, 52.3% did not report the baseline characteristics of their experimental samples: a) whether the control and experimental group were littermates, b) age and c) gender of the animal subjects. The poor reporting of experimental details also hindered subgroup analysis. Lastly, we were bound by the amount of information available, there might be more data available which was never published. There are indications that preclinical studies overestimate the treatment effectiveness by 30% partly due to the absence of published neutral results (*i.e.*, non-significant) and lack of methodological rigor (Sena et al., 2014). A publication bias analysis indicated missing data in multiple categories, and trim and fill analysis showed that the effects on acoustic startle and repetitive behaviour were no longer significant after imputing missing studies, highlighting the consequences publication bias can have.

Taken together, this systematic review and meta-analysis show that the robustness as well as translatability of the *Fmr1* KO model to the clinical profile varies over the different behavioural phenotypes. Overall, many significant phenotypes were found with the same effect direction as seen in patients, thus showing good translational validity. However, altogether there was a large heterogeneity between studies and many effect sizes were relatively small. For most phenotypes there was low replicability which, despite translational validity, asks for careful interpretation of individual study findings. Additionally, when designing a study where the use of the *Fmr1* KO model is considered, one should be aware of the not fully understood mismatch in rodent and clinical phenotypes (*e.g.*, anxiety, startle and PPI, aggression, attention and cognitive flexibility). The cognitive and audiogenic seizure phenotypes showed the highest replicability, in addition to translational validity, therefore the intellectual disability and epilepsy elements of FXS are possibly the most meaningful to study with the *Fmr1* KO. The model as a whole should be more cautiously used for the ASD-like elements of the disorder, which showed translational validity, but the replicability of these phenotypes was low. More importantly, the phenotypic and quality results provided by this meta-analysis urge for a broad reappraisal of the current research and reporting practices in all preclinical models of brain disorders to deliver more meaningful preclinical data.

Acknowledgements

This project has received funding from a ZonMW TOP subsidy (grant number 91216021, 2017) and a MKMD subsidy (grant number 114024154, 2020). In addition, this project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Conflict of interest

The authors have no conflict of interest to declare.

Author contributions

R. Kat: Conceptualization, funding acquisition, investigation, formal analysis, visualization, writing - original draft; **M. Arroyo-Araujo**: Investigation, formal analysis, visualization, writing - original draft; **R. de Vries**: Conceptualization, supervision, writing - reviewing and editing; **M.A. Koopmans**: Investigation; **S.F. de Boer**: Conceptualization, supervision, writing - reviewing and editing; **M.J.H. Kas**: Conceptualization, supervision, funding acquisition, writing - reviewing and editing.

All supplementary files can be found in the online version of the paper

https://www.sciencedirect.com/science/article/pii/S0149763422002111?via%3Dihub

Supplementary file 1 – Search string

Supplementary file 2 – Tests and outcomes table

Supplementary file 3 – Characteristics table

Supplementary file 4 – Risk of bias assessment

Supplementary file 5 – Forest plots

Supplementary file 6 – Publication bias funnel plots

Supplementary file 7 – Sensitivity analysis

Supplementary file 8 – Extracted data

Supplementary file 9 – Meta-analysis results

Supplementary file 10 – Subgroup statistics

Supplementary file 11 – Publication bias analysis statistics

Supplementary file 12 – PRISMA checklist