D4.2 With input from WPs 2 and 3 develop harmonised protocols for cross validation testing

EQIPD – GA 777364
IMI2-2016-09-03

European Quality In Preclinical Data

WP4 – Prospective cross-site validation of guiding principles in specific assay

Lead contributors	Martien Kas (7 – RUG) and Sylvie Ramboz (25 – PGI)
	m.j.h.kas@rug.nl and sylvie.ramboz@psychogenics.com
Other contributors	Yulia Mordashova, Helga Lorenz and Alfred Stefan (20 - AbbVie)
	Heidrun Potschka, Ann-Marie Waldron, and Isabel Seiffert (12 – LMU)
	Janet Nicholson, Heike Schauerte, Tim Ahuis and Patrizia Vöhringer (21 – BI)
	Maarten Loos (14 – Sylics)
	Hanno Wuerbel and Bernhard Voelkl (8 – UBERN)
	Bettina Platt (6 – UNIABDN)
	Tom Van De Casteele (19 – Janssen)

	Maria Arroyo-Araujo (7 – RUG)
	Mathias Jucker and Bettina Wegenast-Braun (EKUT)

Due date	07 May 2019
Delivery date	30 April 2019
Deliverable type	R
Dissemination level	PU

Description of Work	**Version**	**Date**
	V2.0	28 January 2019

Document History

Version	**Date**	**Description**
V1.0	12 April 2019	First Draft
V2.0	23 April 2019	Comments
V3.0	30 April 2019	Final Review
V4.0	07 May 2019	Submission Final Version

Publishable Summary

Following data collection and analysis of the localisation stage experiments (WP4), historical data analysis (WP2), and systematic review of the literature (WP3), harmonised protocols for the three paradigms in WP4 (Irwin test, open field test, and EEG recordings) have been generated. For each of the three paradigms, and based on the guidance of WP2, an inventory of experimental and (laboratory) environmental factors has been collected from each of the partners involved in the prospective studies. Each of the sites has been able to indicate which of these variables could be harmonised in view of feasibility, and those factors will be harmonised as part of the harmonisation protocol. Following harmonisation through standardisation of these factors, the Irwin test and EEG baseline recording experiments from the localisation stage will be repeated according to this standardised protocol. In this way we will test the first hypothesis that harmonisation through standardisation will reduce the between laboratory variability in effect size (when compared to the outcomes from the data obtained during the localisation stage). For the open field studies, harmonisation through heterogenisation versus harmonisation through standardisation will be compared to test the second hypothesis that harmonisation through heterogenisation further reduces the between laboratory variability in effect size (when compared to harmonisation through standardisation and to the localisation stage). A statistical analysis plan has been generated to test these hypotheses. In addition, WP3 has, based on their systematic review and Delphi process, generated a final version of the framework of key principles and criteria for guiding the design, conduct and analysis of preclinical efficacy and safety research. As part of the harmonisation, all sites will adhere to these key principles during the implementation and performance of the harmonised protocol.

Methods

During the localisation stage, data for three paradigms have been collected, namely for the Irwin test (PGI, LMU, BI, ORION, AbbVie) relevant to safety studies, an open field test to assess motor activity levels following an acute pharmacological challenge (RUG, PGI, Sylics, ORION, LMU, UBERN), and a 48 hours of baseline EEG recordings in Wild type and Tg4510 transgenic model for Alzheimer's disease pathology (RUG, PGI, UNIABDN, LMU, BI). Data from these experiments will be described in Delivery report 4.3.

Based on literature studies and insights from historical data collection, WP2 has generated an inventory list with experimental and (laboratory) environmental factors that may influence between laboratory variability (appendix 1). All sites involved in the prospective studies of WP4 have completed these inventories for each of the paradigms. The EEG sites have added a number of extra factors to this inventory, considering that the EEG experiments include a surgical step, as well as complex data analysis (including filtering and artefact removal steps) (appendix 2). All WP4 partners involved in the prospective studies have indicated in the inventory list which of these factors identified can be aligned for the harmonisation protocol (in view of feasibility). Factors that all sites are able to align have been selected and will be standardized as part of the harmonisation protocol.

WP3 has, based on a systematic review and Delphi process, generated a final version of the framework of key principles and criteria for guiding the design, conduct and analysis of preclinical efficacy and safety research (appendix 3). As part of the harmonisation, all sites will adhere to these key principles during the implementation and performance of the harmonised protocol.

In close collaboration with WP2, a statistical analysis plan has been generated to determine whether the between laboratory variability in effect size will reduce following harmonisation through standardisation (and when compared to the localisation stage outcomes). In addition, the sites involved in the open study will test an additional hypothesis, namely whether harmonisation through heterogenisation provides a stronger reduction in between laboratory variability in effect size when compared to the localisation stage and when compared to harmonisation through standardisation.

Results

1. Harmonised protocol for Irwin test

Based on the WP2 and WP4 generated inventories (see appendices 1 and 2), the following factors have been selected for standardisation during the harmonisation protocol for the Irwin study. All Irwin sites involved have indicated feasibility for those factors in view of local implementation possibilities (Table 1). Note that these factors are in addition to the earlier study protocol for the localisation stage study that described the minimal requirements for this study (Milestone report 20). In addition, all sites will adhere to the WP3 derived key principles and criteria for guiding the design, conduct and analysis of preclinical efficacy and safety research (appendix 3).

Table 1. Overview of selected factors for harmonisation through standardisation for the Irwin test.

VarName	VarValue
AgeStart	8.5|8 weeks
AnimalSource	Either Taconic or Charles river, Germany
AnimalCharacteristicsUsedToBalanceGroups	body weight
DrugNaive	yes
DurationOfAcclimatisationToTestRoom	One hour pre-dosing
EnvironmentalEnrichmentType	Yes, enrichment, but source(s) can vary among sites
ExAntePrimaryOutcomeDefined	Yes as per partner site method
ExAnteStatsAnalysisPlan	Yes
ExperimentEnvironment	not home cage
ExperimenterHandlingMethod	tail handling with gloved hands|gloves
FoodRestrictedDuringExperiment	No food available during the assessment time point
HandlingHabituationFreq	2 times per week
HousedInIntervalBetweenAdministrationAndTest	home cage
NumberOfAnimalHandlers	Single handler
NumberofCaretakersInteractingWithAnimals	Multiple
OutcomesPrespecifiedInProtocol	Yes
ParadigmNaive	Yes
PeriodOfDosing	As agree observation time points
PreDefinedHumaneEndpoints	Yes
PreStudyDrinkRestriction	No
PreStudyFoodRestriction	Not applicable for this paradigms
PreviousProcedures	None
RandomisationToOrderAnimalTestingMethod	Example of propose method https://www.randomizer.org/
RandomisationToTestGroupMethod	Yes
ResearcherPresentDuringTestPhase	Yes
RouteOfAdministration	intraperitoneal
Sex	Both
SocialHousingStandardisation	Yes

VarName	VarValue
StandardisationOfOutcomeAssessment	observations by experimenter trained on a fixed assessment protocol (sequence of actions; handlings)
StudyAsPerExAnteProtocol	Yes
TestAreaCleanedBetweenAnimals	Yes
TestAreaCleaningMethod	ethanol|Not Applicable
TestArenaBeddingType	None
WasExperimenterBlindDuringTestPhase	Yes
WasHandlerBlinded	Yes
WasHandlingStadardised	Yes
WaterRestrictionDuringExperiment	No water available during the assessment time point
CageStandardisation	Yes
CaretakerHandlingMethod	tail handling with gloved hands|gloves
CleaningFrequencyPerWeek	1
Drink	tap water
DrinkAccess	ad libitum
ResearcherASmoker	No
WaterFoodReplacementFreq	Weekly
BeddingTransfer	no|bedding
ConsistentNumberPerCage	Yes if no separation needed due to animals fighting

In addition to the environmental factor harmonisation protocol, the Irwin test will also include a harmonisation in the scoring method for each outcome variable measured enabling findings to be compared between sites. As an example the scoring method could consist in attributing a zero “0” value to a normal response, state or absence of, and an increment numerical value plateauing to four “4” indicated the deviation or severity from normal response or animal state. The total of the outcome variable measures from each site will not be altered.

2. Harmonised protocol for open field test

Based on the WP2 and WP4 generated inventories (see appendices 1 and 2), the following factors have been selected for standardisation during the harmonisation protocol for the open field study. All sites involved in the open field study have indicated feasibility for those factors in view of local implementation possibilities (Table 2). Note that these factors are in addition to the earlier study protocol for the localisation stage study that described the minimal requirements for this study (Milestone report 20). In addition, all sites will adhere to the WP3 derived key principles and criteria for guiding the design, conduct and analysis of preclinical efficacy and safety research (appendix 3).

In addition to the harmonisation protocol for the Irwin test and EEG study, the open field study will include not only a harmonisation through standardisation procedure, it will also include a repetition of the localisation stage, as well as a harmonisation through heterogenisation procedure. The repetition of the localisation stage is possible for the open field test and will be implemented to test internal reproducibility, and to generate data in parallel with the harmonisation approaches for direct comparison. For these experiments, only female C57BL/6J mice will be tested; and only 1 experimental condition (vehicle versus MK801 treatment (0.2 mg/kg) will be tested. There are two reasons to only select one sex for the harmonisation study, namely, 1) both males and females showed the same responses during the localisation stage, 2) since we are also introducing harmonisation through heterogenisation, we need to reduce the number of experimental groups to account for the number of animals needed for this extra harmonisation condition (i.e., to stay within the number of animals applied for in the ethical approvals). For the localisation repeat, the exact same protocol will be performed as has been done previously for the localisation stage experiment. For the harmonisation through standardisation protocol, the localisation stage protocol (see above) will be performed, however, the factors indicated in the table below will be standardised for each site. For the harmonisation through heterogenisation protocol, the harmonisation through standardisation protocol will be performed, however, one factors will be systematically changed in half of the animals tested for this protocol. This factor is the light intensity of the testing arena, in which half of the animals will be tested at 20 lux and half of the animals will be tested at 250 lux. For the harmonisation through standardisation, animals will be tested at 50 lux.

Table 2. Overview of selected factors for harmonisation through standardisation and for harmonisation through heterogenisation for the open field test.

VarName	VarValue
AgeStart	9-10 weeks
AllOutcomesReported	Yes, according to agreed endpoints
AnimalCharacteristicsUsedToBalanceGroups	Sex
AutomationMethod	Video tracking and analysis
QC of Data	Tracking was verified by experimenter. Based on these inspections, specific trials were edited within Ethovision such that each point was accurately scored and issues associated with automated tracking were eliminated. The detection settings for tracking were selected so that both the percentage of samples in which the subject was not found, and the percentage of samples skipped were less than 1% per trial
RandomisationToOrderAnimalTestingMethod	Random number generator, Example of propose method https://www.randomizer.org/
RandomisationToTestGroupMethod	Random number generator, Example of propose method https://www.randomizer.org/
TestArenaLightIntensity	**50 lux (for harmonisation through standardization); 20 and 250 lux (for harmonization through heterogenisation)**
WasExperimenterBlindDuringTestPhase	Yes
WasHandlerBlinded	Yes

3. Harmonised protocol for baseline EEG recordings

Based on the WP2 and WP4 generated inventories (see appendices 1 and 2), the following factors have been selected for standardisation during the harmonisation protocol for the EEG study. All EEG sites involved have indicated feasibility for those factors in view of local implementation possibilities (Table 3). Note that these factors are in addition to the earlier study protocol for the localisation stage study that described the minimal requirements for this study (Milestone report 20). In addition, after the study, all sites involved will collect tissue to genotype the animals used for the harmonisation study in order to make sure that genotypes are matching with the expected genotypes. In addition, all sites will adhere to the WP3 derived key principles and criteria for guiding the design, conduct and analysis of preclinical efficacy and safety research (appendix 3).

Table 3. Overview of selected factors for harmonisation through standardisation for the EEG baseline recordings.

VarName	VarValue
(Bregma coordinates)	AP -2mm; ML -1.5mm
Parameters analyzed	raw and relative power
AgeStart (e.g. 10 weeks)	22-23 weeks olds at the start of the experiment
SocialHousingStandardisation	Before surgery: group housed, 4 animals per cage, Tg housed with Tg, WT housed with WT.

VarName	VarValue
EnvironmentalEnrichmentStandardisation	Yes, enrichment, but source(s) can vary among sites
ConsistentNumberPerCage	After surgery, yes, 2-3 animals per cage, genotype matched.
HandlingHabituationFreq	Yes, 2 times per week
StartDateTimeOfTestPhase (HH:MM:SS)	4th hour into light cycle
StudyDesign	completely randomized
AnimalCharacteristicsUsedToBalanceGroups	Genotype
RandomisationToOrderAnimalTestingMethod	Random number generator, Example of propose method https://www.randomizer.org/
WasExperimenterBlindDuringTestPhase	Yes
WasHandlerBlinded	Yes
WasHandlingStandardised	Yes, tail handling with gloved hands
HandlingMethodDuringStudy	Tail handling with gloved hands
NumberOfOutcomesMeasured (e.g. 10)	EEG power spectrum according to agreed data endpoint sheet
StudyAsPerExAnteProtocol	Yes, experimental procedures are executed according to an internally approved ex ante study protocol
OutcomesPrespecifiedInProtocol	Yes
ExAnteStatsAnalysisPlan	Yes, the outcomes to be measured were described in an ex ante study protocol.
AllOutcomesReported	No, only EEG power spectrum according to agreed data endpoint sheet
QC of Data	Yes, details will be discussed later
Was experimenter handling the animals blind to test group allocation?	Yes
Location of recording electrodes	At the level of Hippocampus (skull electrode)
Location of reference electrodes	At the level of Cerebellum (skull electrode)

4. Statistical analysis plan

(Generalized) linear (mixed) models will be used to estimate and test effects of treatment on primary outcomes for each of the three considered paradigms, harmonisation conditions (localised, standardised and/or heterogenized) and experimental designs. Outcome variables will be transformed as appropriate if distributional assumptions of the outcome data are not fulfilled or to ensure numerical convergence of computational algorithms. For theoretical statistical background we refer to Verbeke and Molenberghs (2000), Fitzmaurice et al (2004) and Molenberghs and Verbeke (2005). All significance tests are two-tailed with a type I error of 0.05.

1. Analysis by laboratory and condition

To evaluate the reproducibility of significance testing of scientific hypotheses of interest in the way research results are typically reported in scientific literature, data generated by the laboratories under different conditions of harmonisation, will be analysed separately. Linear models will be used to estimate group means and to compare outcomes between treatment groups. Differences between treatment groups in mean primary outcomes will be tested for statistical significance based on two-tailed Wald-tests with a type I error of 0.05.

2. Analysis across laboratories and conditions

To study sources of (lack of) reproducibility across laboratories and conditions, two statistical approaches will be used.

As a primary analysis, variation in treatment effects across laboratories and conditions will be estimated with a (general) linear model, with treatment, laboratory, conditions and relevant interactions as fixed factors. F-test with Kenward-Roger adjusted degrees of freedom will be used to test signifance of effects. Significance tests for contrasts (e.g. for pairwise group comparisons) will be based on least-square means. See also Wolfinger

(2013) for an analysis of reproducibility of animal experiments using similar methodology for slightly different experimental designs.

As secondary analyses, variation in treatment effects across laboratories and conditions will be estimated and tested with (generalized) linear mixed models using restricted maximum likelihood (REML), with treatment (and/or conditions and their interactions) as fixed factors, and random factors to estimate sources of variation for treatment effects at the level of laboratories and conditions. Variance-covariance structure for random effects and for residual errors will selected based on a model building exercise as explained in Verbeke and Molenberghs (2000), Fitzmaurice at el (2004) and Molenberghs and Verbeke (2005). In the case that the number of studied laboratories and sample sizes per laboratory are too limited to draw reliable inference using the methodologies references above, appropriate resampling techniques (Fitzmaurice et al. 2007, Drikvandi et al. 2013) will be applied to test variance components for significance, and/or to compare variance components between conditions.

Conclusion

For the harmonisation stage, for all three paradigms harmonisation protocols have been established based on the information from WP2, WP3, and WP4. Following the indicated statistical analysis plan, it will now be tested whether harmonisation through standardisation reduces between laboratory variability in effect size for the Irwin test, open field test, and EEG baseline paradigms (when compared to localisation stage outcomes). In addition, for the open field test, it will be tested whether harmonisation through heterogenisation reduces between laboratory variability in effect size further when compared to harmonisation through standardisation.

Repository for primary data

All raw and endpoint data for the localisation stage and the harmonisation stage, as well as the information from all sites obtained following completion of the inventories will be uploaded in the EQIPD central data base.

References

Drikvandi, R., Verbeke, G., Khodadadi, A. and Partovina, V. (2013) Testing multiple variance components inn linear mixed-effects models. Biostatistics 14, 144-159.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004) Applied longitudinal analysis. Wiley: New Jersey.

Fitmaurice, G.M., Lipsitz, S.R. and Ibrahim, J.G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics 63, 942-946.

Molenberghs, G. and Verbeke, G. (2005). Models for discrete longitudinal data. Berlin: Springer.

Verbeke, G. and Molenberghs, G. (2000). Linear mixed models for longitudinal data. Berlin: Springer.

Wolfinger (2013). Reanalysis of Richter et al. (2010) on reproducibility. Nature Methods 10, 373-374.

Appendix 1: Inventory lists of potential experimental and (laboratory) environmental factors for harmonisation

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
EnvFactorsPrestudy	1. Husbandry conditions	What type of food was given to the animals?		pellet, powder, …	StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	What type of food dispenser was used?			StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	What was the access schedule to food (over 24 hours)?		ad libitum, restricted	StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	Was animal on food restriction just before the experiment?		yes, no	AnimalID/TestgroupID/StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	What type of drink was given to the animals?		tap water, ...	StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	How often was water/drink replaced?		daily, …	StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	What was the access schedule to drink (over 24 hours)?		ad libitum, restricted	StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	Was animal on water restriction just before the experiment?		yes, no	AnimalID/TestgroupID/StudyID	
EnvFactorsPrestudy	1. Husbandry conditions	Were husbandry conditions standardized across studies?		yes, no	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	Were housing conditions of the home cage standardized across studies with respect to caging, noise, social housing, animal management/ma		yes, no	StudyID	see Toth 2015

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
		nipulation and environmental enrichment?				
EnvFactorsPrestudy	2. Housing conditions home cage	What was light/dark schedule per 24 hours in home cage?		e.g. 12L-12D, 6L-6D-6L-6D, …	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	What was light intensity during light phase in home cage?	Lx	[exact value or range]	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	What was humidity in home cage?	%	[exact value or range]	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	How many animals were housed per home cage?			StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	Did all cages have the same number of animals?		yes, no	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	How many care-takers interacted with the animal before the experiment?			StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	How were animals handled by care takers?		bare hands, gloves, tail handling, cupped hands, transfer box/tunnel, net, forceps, other	StudyID	Gouveia and Hurst (2017)
EnvFactorsPrestudy	2. Housing conditions home cage	What kind of environmental enrichment was used in home cage?		none, sizzling material for nest building, tubes, objects for climbing, wood for biting,…	StudyID	
EnvFactorsPrestudy	2. Housing conditions	What was the bedding material of the home		sawdust, shredded paper, pellet,	StudyID	

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
	home cage	cage?		aspen-chip, corncob, SaniChips, other		
EnvFactorsPrestudy	2. Housing conditions home cage	How many times per week were home cages cleaned?			StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	Was a little of the dirty bedding transferred into clean cage to help the animals settle?		yes, no	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	What was home cage temperature?	Degrees Celsius	[exact or range]	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	What means of ventilation were used?		none, passive (e.g. open window, …), ventilation system without temperature control, ventilation system with temperature control, …	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	Was the facility where the home cages are located maintained at SPF		yes, no	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	If the facility where the home cages are located, was maintained at SPF, specify level		SPF 2014, SOPF 2014, SPF 2012, SOPF 2012, other	StudyID	
EnvFactorsPrestudy	2. Housing conditions home cage	What was the shape of the home cage?		square, circular, …	StudyID	
EnvFactorsPrestudy	2. Housing conditions	What was the size of the home cage (square:	Cm		StudyID	

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
	home cage	width x depth x height; circular: diameter x height, ...)?				
EnvFactorsPrestudy	2. Housing conditions home cage	With respect to animal's home cage shelf level, did animal(s) have neighbours above, next to or under their cage?		no, above shelf, next to, under shelf, …	AnimalID/StudyID	next to = on equal height; list all that apply
EnvFactorsPrestudy	3. Transportation from home cage to test arena	How were animals transported from home cage to test arena?		trolley, carried cage, …	StudyID	
EnvFactorsPrestudy	3. Transportation from home cage to test arena	What measures were taken to limit influence on animal stress levels of the transport from home cage to test arena (visual, sound, …)?		none, covered cage, …	StudyID	
EnvFactors Prestudy	3. Transportation from home cage to test arena	Where was the animal housed in between administration and start of the test?		home cage, test apparatus, …	StudyID	

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
EnvFactorsStudy	1. Experimenter-animal interaction	How are animals handled by experimenters?		bare hands, gloves, tail handling, cupped hands, transfer box/tunnel, net, forceps, other	StudyID	Gouveia and Hurst (2017)

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
EnvFactorsStudy	1. Experimenter-animal interaction	Were animals habituated to the handling method?		yes, no	StudyID	
EnvFactorsStudy	1. Experimenter-animal interaction	What was researcher's sex?		male, female, both, unknown	StudyID	list all in case there were multiple experimenters
EnvFactorsStudy	1. Experimenter-animal interaction	What intensive smell(s) did researcher carry?		none, perfume, garlic, unknown, …	StudyID	
EnvFactorsStudy	1. Experimenter-animal interaction	Was researcher a smoker?		yes, no	StudyID	
EnvFactorsStudy	1. Experimenter-animal interaction	Was researcher experienced or inexperienced with paradigm?		experienced, inexperienced but trained, inexperienced and not trained, …	StudyID	
EnvFactorsStudy	1. Experimenter-animal interaction	Was researcher in the room of the test arena during the test phase?		yes, no, visually hidden	StudyID	
EnvFactorsStudy	1. Experimenter-animal interaction	How many experimenters interact with the animal?		single, multiple	StudyID	
EnvFactorsStudy	2. Animal preparation	How long were animals left to acclimatise and habituate to testing area before the experiment?		hours, or NA in case housing and testing room were the same	StudyID	
EnvFactorsStudy	2. Animal preparati	What was duration of surgery?	Hours		StudyID	

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
	on					
EnvFactorsStudy	2. Animal preparation	How long before start of the experiment was surgery performed?	mins		StudyID	
EnvFactorsStudy	2. Animal preparation	What method of anaesthesia/analgesia was used?			StudyID	
EnvFactorsStudy	2. Animal preparation	What anaesthetic was used?		%O2, %NO2, medical air?, …	StudyID	
EnvFactorsStudy	2. Animal preparation	How long were animals fasted prior to the anaesthetic routine?	Hrs		StudyID	
EnvFactorsStudy	2. Animal preparation	Were animals orally intubated and artificially ventilated during the anaesthetic routine?			StudyID	
EnvFactorsStudy	2. Animal preparation	What other support do animals receive during anaesthesia?		depth of anaesthesia controlled, etc.	StudyID	
EnvFactorsStudy	2. Animal preparation	What surgical procedure was used?			StudyID	
EnvFactorsStudy	2. Animal preparation	Was fluid/saline given during surgery?		yes, no	StudyID	
EnvFactorsStudy	2. Animal preparation	Was temperature of the animal during surgery kept under control?		yes, no	StudyID	e.g. heating material or thermometer?
EnvFactorsStudy	3. Physical characteristics test arena	Where (physical location) was the experiment executed?	[city (country)]		StudyID	name city and country
EnvFactorsStudy	3. Physical characte	What was the shape of the test arena?		square, circular, …	StudyID	Could be used to scale outcome measures across

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
	ristics test arena					studies/contributors
EnvFactorsStudy	3. Physical characteristics test arena	What was the size of the test arena (square: width x depth x height; circular: diameter x height; ...)?	Cm		StudyID	Could be used to scale outcome measures across studies/contributors
EnvFactorsStudy	3. Physical characteristics test arena	What was the color of the walls of the test arena?		black, grey, white, …	StudyID	
EnvFactorsStudy	3. Physical characteristics test arena	What was the bedding material of the test arena?		sawdust, shredded paper, pellet, aspen-chip, corncob, SaniChips, other	StudyID	
EnvFactorsStudy	3. Physical characteristics test arena	How many test arenas were available for use?		one, two, more than two	StudyID	
EnvFactorsStudy	3. Physical characteristics test arena	What was the physical environment of experiment?		home cage, environment different from home cage, …	StudyID	
EnvFactorsStudy	3. Physical characteristics test arena	Was test arena cleaned between testing of different animals?		yes, no	StudyID	
EnvFactorsStudy	3. Physical characteristics test arena	If test arena was cleaned between testing of different animals, what cleaning material was used?		ethanol, disinfectant wipes, …	StudyID	
EnvFactorsStudy	4. Social conditions during test	How many animals were tested at the same time in the		one, two, more than two	StudyID	

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
	phase	same test arena?				
EnvFactorsStudy	5. Temporal characteristics test phase	What was date-time of (start of) the test phase of the experiment?	datetimeformat		AnimalID/StudyID	
EnvFactorsStudy	5. Temporal characteristics test phase	What was the time of (start of) the test phase of the experiment relative to day/night cycle?	Mins	x mins after start of night phase	AnimalID/StudyID	
EnvFactorsStudy	5. Temporal characteristics test phase	What was the total duration of the test phase?	Mins		StudyID	Could be used to scale outcome measures across studies/contributors
EnvFactorsStudy	6. Food/drinking conditions	Was animal on food restriction during the experiment?		yes, no	AnimalID/TestgroupID/StudyID	
EnvFactorsStudy	6. Food/drinking conditions	Was animal on water restriction during the experiment?		yes, no	AnimalID/TestgroupID/StudyID	

Domain	**Subcategory source of data heterogeneity**	**HeterogenSource_varname**	**HeterogenSource_varunits**	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
ExpDesign	1. Test groups	Type of effect studied		pharmacological effect, genotypic effect, other	StudyID	
ExpDesign	1. Test groups	Type of study design		tbd (training)	StudyID	check Festing and Altman 2002 and http://www.3rs-reduction.co.uk/html/9__experimental_designs.html for training: completely randomized, randomized block (“within-subject”, “crossover”, “matched subjects”), factorial, Latin square,

						crossover, repeated measures, split-plot, incomplete block, sequential
ExpDesign	1. Test groups	What test groups/factors/sequences were included in the study (only consider study test phases)?		tbd (training) e.g. vehicle, test_cpd_xxx–dose, …	StudyID	Indicate ALL test groups of the study, not only the test groups that contain non-proprietary information
ExpDesign	2. Dosing regimens	If pharmacological intervention was applied, what was the dosing regimen of drug administration?		acute, chronic	StudyID	
ExpDesign	2. Dosing regimens	If dosing regimen was chronic, what was the frequency of drug administration?		once daily, …	StudyID	
ExpDesign	2. Dosing regimens	If dosing regimen was chronic, what was the duration of drug administration (in days)?			StudyID	
ExpDesign	2. Dosing regimen	If pharmacological intervention was applied, what was the time of (first) drug administration relative to start of the test phase?			StudyID	
ExpDesign	3. Drug	If pharmacological intervention was applied, what drug was used?		test compound name, vehicle name, negative control name, positive control name	TestgroupID	include whether free base or salt was used
ExpDesign	3. Drug	If pharmacological intervention was applied, what drug formulation was used?		liquid, …	TestgroupID	
ExpDesign	3. Drug	If pharmacological intervention was applied, what drug dose was used?		[in mg/kg]	TestgroupID	
ExpDesign	3. Drug	If pharmacological intervention was applied, what was the route/site of drug administration?		tail vein, femoral vein, jugular vein, subcutaneuous, intracerebroventricular, osmotic minipump,	TestgroupID	

				oral gavage, intraperitoneal, other		
ExpDesign	3. Drug	If pharmacological intervention was applied, what type of solvent was used?		DMSO, PBS, other	TestgroupID	
ExpDesign	3. Drug	If pharmacological intervention was applied, what solvent dose was used?		[in mL/kg]	TestgroupID	
ExpDesign	4. Genotype	If the study evaluated a genotypic effect, specify the genotype		WT, TG (HO, HET), …	TestgroupID	

Domain	**Subcategory source of data heterogeneity**	HeterogenSource_varname	HeterogenSource_varunits	**HeterogenSource_value**	**Data level**	**Argumentation/reference/comment**
InternValid	1. Randomisation	What randomisation method was used for allocation of animals to test groups?		unknown, none, flip a coin, shuffle sealed envelope, random number generator, alternate allocation, day of the week, odd versus even, pick randomly from cage, other	StudyID	
InternValid	1. Randomisation	For what animal characteristics was allocation to test groups matched or balanced?		unknown, none, body weight, sex, age, genotype, other	StudyID	
InternValid	1. Randomisation	What randomisation method was used for randomisation of the order of testing of animals across test groups?		unknown, none, flip a coin, shuffle sealed envelope, random number generator, alternate allocation, day of the week, odd versus even, pick randomly from cage, other	StudyID	i.e. to avoid systematic difference between test groups in daytime (e.g. morning vs afternoon), etc …

InternValid	2. Optimized experimental design	Was pharmacokinetics or dose-response information used to select drug dose for treatment?		yes, no	StudyID	Hypothesized effect on bias/precision of data through research conduct is unclear; dose can be used as a covariate (experimental design factor), independently from research conduct
InternValid	2. Optimized experimental design	What sample size rationale was used for the study?		none, from historical experiments, from literature, formal sample size calculation based on knowledge of variability and assumed effect size	StudyID	
InternValid	3. Blinding	Was experimenter blind to test group allocation during the test phase?		yes, no	StudyID	
InternValid	3. Blinding	Was experimenter handling the animals blind to test group allocation?		yes, no	StudyID	
InternValid	4. Prespecification/standardisation of experimental procedures	Was handling of animals standardized (protocol description and training) across experimenters (and studies)?		yes, no	StudyID	
InternValid	4. Prespecification/standardisation of experimental procedures	What handling method was used?		tail, cupped hand, tunnel	StudyID	Gouveia and Hurst (2017)
InternValid	4. Prespecification/standardisation of experimental procedures	How many outcomes were measured?			StudyID	

InternValid	4. Prespecification/standardisation of experimental procedures	Were experimental procedures (up to and including data collection) executed according to an internally approved ex ante study protocol?		yes, no	StudyID	
InternValid	4. Prespecification/standardisation of experimental procedures	Were the outcomes to be measured described in an ex ante study protocol?		yes, no	StudyID	
InternValid	4. Prespecification/standardisation of experimental proceduress	How was outcome assessment standardized?		observations by experimenter without explicit training on a fixed assessment protocol (sequence of actions, handlings), observations by experimenter trained on a fixed assessment protocol (sequence of actions, handlings), automated assessment	StudyID	
InternValid	4. Prespecification/standardisation of experimental procedures	If outcome assessment was done in an automated way, what type of automation was used?		Free field: name of technology and software	StudyID	
InternValid	4. Prespecification/standardisation of experimental procedures	Were statistical analyses described in an ex ante statistical analysis plan?		yes, no	StudyID	Not relevant as a factor for meta-analysis; only relevant for aggregate data
InternValid	4. Prespecification/standardisation of experimental procedures	Was primary outcome measure defined in an ex ante study protocol?		yes, no	StudyID	
InternValid	4. Prespecification/standardisation of experimental procedures	What were pre-defined criteria for withdrawal of animals (humane endpoints) from			StudyID	

		the study before the primary endpoint was reached?				
InternValid	5. Data selection	Was outcome reported for all animals that were assigned to test groups (including missing observations)?		yes, no	StudyID	
InternValid	6. Data quality control	Were data qc-ed (e.g. checked for reporting errors, …)		yes, no	StudyID	

Appendix 2: Additional inventory lists of potential experimental and (laboratory) environmental factors for harmonisation of EEG baseline recording studies.

Tethered or telemetry	
Screw size or array weight/size	
Number of recording electrodes	
Location of recording electrodes	
(Bregma coordinates)	
Number of reference electrodes	
Location of reference electrodes	
Cementing agent	
Anesthesia	
Dose	
Route of administration	
Length of time under anesthesia	
Local anesthesia	

Analgesia	
Dose	
Route of administration	
Frequency	
Recording software	
Analysis software	
Parameters analyzed	
Sampling Rate	
Sampling Filters	
Sampling Gain	

Appendix 3: Key principles and criteria for guiding the design, conduct and analysis of preclinical efficacy and safety research

How to read this appendix

The following is the framing of the result of a systematic review of existing guidelines for reducing bias in preclinical research, two subsequent rounds of Delphi in the consortium, and a consensus meeting, agreeing on the 33 items that can be found at the end of this document. These can be reidentified in the text via the [numbers in brackets]. Specific terms where, whenever possible, brought in line with the ARRIVE guidelines for reporting of preclinical experiments and the Experimental Design Assistant of the United Kingdom's National Centre for the 3Rs, and the PREPARE guidelines for designing preclinical experiments of Norway's National Consensus Platform for the advancement of the 3Rs.

A guiding rope, not a ligature around the neck

While we consider the domains listed below as generally applicable to all preclinical experiments, specific items may be challenging or impractical in some settings. They are meant to set examples for the spirit of the domains, but they are neither applicable to every experimental setting, nor are they a comprehensive list. Instead, use them as food for thought to adapt to your situation. If you choose to not follow any of the items below, provide transparency by documenting this, including your justification for not doing so.

DOMAIN 1: exploratory vs. confirmatory research

"Answer the following question: Is your experiment testing a predefined scientific hypothesis which is statistically testable (confirmatory research) or is it exploring a space of interesting options to generate hypotheses (exploratory research)?"

When planning an experiment, first consider whether you aim to perform a confirmatory experiment or an exploratory experiment. A general problem in scientific publications of the previous decades is suspected to be the large number of exploratory experiments that have been effectively treated and reported as confirmatory, which led to an overestimation of effect sizes [29].

- There is a one-way street between confirmatory and exploratory experiments: if you find interesting results that go against your hypothesis, a confirmatory experiment can turn into an exploratory experiment. An extensive study can include confirmatory experiments (one clear hypothesis and a fitting study design to test it in a well-powered population), as well as exploratory experiments or analyses (additional questions to investigate out of curiosity). However, an exploratory experiment can never become confirmatory: the result of an exploratory experiment or analysis should never be disclosed, disseminated or published as driven by hypothesis. This is not only a question of transparency (hypothesizing after the results are known (HARKing) while disseminating as pre-specified hypothesis), it also affects the false discovery rate (i.e. the chance that random findings appear significant).

- Any confirmatory experiment will at least need: a clear, predefined hypothesis and a clear primary outcome measure to test the hypothesis in an appropriate statistical test. Furthermore, a meaningful predefined sample size is required [6,7,9,10].

- Exploratory experiments face their own challenges: consider not using probability statistics (i.e. anything that produces a p-value) at all, but rather report effect sizes along with a measure of uncertainty (e.g. confidence intervals) [24]. Keep in mind that multiple testing will produce random findings, so be aware of the limited extrapolation capacity of the findings of an exploratory experiment.

DOMAIN 2: pre-planning and standard operating procedures (SOPs)

"Whenever and where ever meaningful, possible and feasible, prespecify, document and standardize all methods and analyses before the experiment."

Standardization can reduce between-study heterogeneity in experimental conduct and can help in making experiments easier to replicate. It also makes experiments more transferable, e.g. into other labs, onto new equipment or to different personnel.

- Determine what is already known – consider performing a systematic search or systematic review, and search databases of registered protocols to avoid unnecessary repetition and inform your planning [25].
- The use of electronic lab books can be an efficient first step in standardized documentation. Consider pre-registering your protocol in a repository, i.e. before the experimental conduct. This will increase transparency, avoid unnecessary duplication of your work by others, and also make it easier to publish your results irrespective of outcome [26]. Some journals offer pre-registered studies, where the protocol of your experiment is reviewed and, if accepted, publication of the results guaranteed – consider this option where feasible [5].

- Include meaningful negative and positive and controls in your experiment [11] (i.e. include an experimental condition/ group where the outcome will certainly be negative, such as baseline measurements, vehicle treatment, or no treatment, and an experimental condition/

group where the outcome will certainly be positive, e.g. a gold-standard of treatment). Include these groups in your hypothesis.

- Define in- and exclusion criteria for allocation of subjects into the experiment / experimental groups [5,15].
- Think about disclosure and a dissemination plan: Keep your full dataset (and consider publishing it) and use appropriate granular ways of informative data display that do not hide outliers (e.g., dot plots rather than bar graphs, overlie boxplots with violin or dot plots).
- Use calibrated instruments, define standardized animal and/or sample handling procedures, define standardized analyses procedures, and train each participating experimenter in these procedures, to reduce inter-experimenter variation [4,31,32,33]. Document the source of reagents.

DOMAIN 3: statistics

"Think about which form of aggregate measures are meaningful for your data and choose appropriate statistical methods."

T-tests and p-values have their "raison d'être" but are overrepresented in scientific papers. Bear in mind that often the data do not follow a normal distribution and therefore violate the assumptions of parametric tests such as t-tests and ANOVAs. Correction for multiplicity will be necessary when performing multiple analyses. Furthermore, for exploratory experiments you should consider not using p-values and hypothesis-testing statistics at all.

- Formulate your statistical analysis plan before the study. This can be a paragraph in your protocol and it needs to consider all (confirmatory and experimental) analyses you plan to conduct. Provide details of all methods planned for each part of the study [6]. Consult a statistician – there is no need to suffer through this alone.
- If your experiment is confirmatory, define a clear hypothesis and a clearly testable outcome measure [6,7]. Base your calculation of sample size on the literature or estimated sample variation and relevant effect size, or if not feasible, on guidelines for sample sizes for non-parametrical tests [10]. Stick to your precalculated sample size. A dynamic change of sample size is not appropriate in close to all experimental designs.
- State all your additional secondary hypotheses and choose adequate statistics and outcome measures to test them. Correct for multiple testing or choose a statistical method that is suited for multiple outcomes [8].
- If your experiment is exploratory, consider using statistical measures of precision rather than probability (e.g., present effect sizes along with confidence intervals rather than p-values [24]). If you want to stick to p-values, choose a method to correct for multiple testing.
- All statistical approaches follow certain assumptions – test if your data meet them. To start with, is your data distributed normally? If it is clearly not, t-test, ANOVA and similar methods are skewed, and you should choose a non-parametric statistical method set [9]. Consider analysis that take e.g. repeated measures, multiple variables (with or without co-linearity) into account, where applicable.
- Think about what your unit of analysis is, what are technical and what independent replicates. For example, every sample coming from the same animal (i.e. multiple cells in an in-vitro experiment or multiple repetitions of a task in an in-vivo experiment) is not

independent and should therefore be averaged per animal before going into comparison [23].

DOMAIN 4: Randomization and blinding

“Whenever and where ever meaningful, possible and feasible, randomize and blind your processes to avoid the introduction of confounding and systematic error.”

The difficulty of independently replicating results from other groups, or even your own results from a month ago, may often be the result of a substantial degree of bias, leading to systematic error that was unintentional or unknowingly introduced to the experiment. In order to optimally assess the causal relationship between an intervention and the outcome, you need to reduce the risk of bias as much as possible.

- Randomize the allocation of animals to experimental groups [3,22]. All groups (including all controls) need to be randomized.
- Use standardized processes or software to randomize rather than chance or pseudo-random procedures. Most haphazard methods may seem random but may actually hold a significant chance of introducing bias [1,3]. Document the method used.
- Conceal the allocation sequence, if feasible. This is the first level of blinding (or masking) and helps to reduce the risk of bias caused by expectations of the experimenters from the start [2].
- Keep your experimenters blinded and the allocation masked at multiple stages, e.g. experimental conduct, outcome assessment and analysis processes. Keeping the experimenter neutral at as many stages as possible will reduce the risk of biased performance or outcome detection [1,30].
- Matching or balancing animals for age group, sex, or variables specifically important for your experiment may be important to help avoid the introduction of confounds to your results. Methods for stratified randomization (that keep balance of key aspects during randomization) can be helpful. In nearly all experiments, you will find that balancing or matching at least age group and sex is mandatory [21], but other aspects will be specific for your setting. Draw an informed opinion from literature about which additional factors you need to address as potential confounders. Usual suspects include type of anesthesia, treatment, (experimental) setting or co-morbidities [17,18,19,20].

DOMAIN 5: documentation

“Not all bias can be avoided, but most can be uncovered: use full and comprehensive documentation.”

You will not be able to foresee everything that influences the outcome of your experiment, and you cannot fully control each variable. Therefore, it is of highest importance that you document everything of general importance and any deviations from the planned protocol, such as unexpected events during your experiment.

- Keep track of your animal characteristics at the beginning of the experiment (baseline) and, where meaningful, also during the experiment. The list of characteristics will differ depending on the nature of the experiment, though some variables are universal (e.g. health status, weight or age and sex should be documented in general). Genetic background and breeding scheme will be of high importance for transgenic animal lines and relevant in other cases [14,27]. Physiological variables might be important to keep track of during the experiments in many cases [12,28]. Animal housing and environmental conditions need to be documented as well.

- Keep track of the flow of each animal or sample through your experiment(s). In case of exclusions or drop-outs, specify which experimental group these animals belonged to and what prespecified exclusion criteria applied to each of them. If none of the predefined exclusion criteria applied, document the alternative reason for exclusion or drop-out [13].
- Keep track of accidental unblinding and other unexpected events you observe during the experimental conduct, as well as any deviation from the protocol. Unforeseen circumstances may require deviations from the planned protocol. However, if you document and clearly communicate these, transparency is kept, enabling you and others to interpret your study results in the light of these deviations. It can also help inform future experimental design [5,16].

List of items generated from systematic review of existing guidelines, two subsequent rounds of Delphi process within the consortium, and final discussions at a consensus meeting

[1] Blinding of outcome assessment and analyses (incl. Method and procedures for unblinding)

[2] Concealed allocation of treatment sequence

[3] Randomized allocation of animals to treatment (incl. Description of the method)

[4] Standardized handling of animals within the experiment (way how they are treated should be defined in the SOP)

[5] Pre-registration of study protocol and analysis procedures incl. Hypotheses and variables and procedures for exclusion of outliers. Description of all deviations from the protocol

[6] A priori statements of hypothesis

[7] Definition of outcome measurement criteria

[8] Establishment (and clear ranking) of primary and secondary end points and outcomes

[9] Choice of adequate statistical methods

[10] Adequate choice of sample size

[11] Selection of appropriate (negative and positive) control groups

[12] Characterization of animal properties at baseline

[13] Recording of the flow of animals through the experiment

[14] Reporting on genetic background

[15] Description of in/exclusion criteria and data censoring

[16] Monitoring emergence of confounding characteristics in animals

[17] Addressing confounds associated with anaesthesia or analgesia

[18] Addressing confounds associated with treatment

[19] Addressing confounds associated with setting or experimental setting

[20] Addressing treatment interactions with clinically relevant co-morbidities

[21] Matching or balancing sex of animals across groups

[22] Matching or balancing treatment allocation of animals

[23] Specification of unit of analysis

[24] Precision of effect size

[25] Critical appraisal of literature or systematic review during design phase

[26] publish negative results

[27] Reporting on breeding scheme

[28] monitoring and regulation of physiological variables

[29] statement of whether the study is hypothesis-testing or hypothesis-generating (confirmatory vs exploratory)

[30] blind performance of experimental procedures

[31] Personnel is adequately trained

[32] Apparatus is calibrated and/or correctly maintained

[33] Sources of reagents