Refinement of Experimental Design and Conduct in Laboratory Animal Research

Jeremy D. Bailoo, Thomas S. Reichlin, and Hanno Würbel

Abstract

The scientific literature of laboratory animal research is replete with papers reporting poor reproducibility of results as well as failure to translate results to clinical trials in humans. This may stem in part from poor experimental design and conduct of animal experiments. Despite widespread recognition of these problems and implementation of guidelines to attenuate them, a review of the literature suggests that experimental design and conduct of laboratory animal research are still in need of refinement. This paper will review and discuss possible sources of biases, highlight advantages and limitations of strategies proposed to alleviate them, and provide a conceptual framework for improving the reproducibility of laboratory animal research.

Key Words: 3R; refinement; ARRIVE; reproducibility; internal validity; external validity; standardization; preregistration

What Is the Problem?

In 2005, the biomedical research community was startled by a paper entitled "Why Most Published Research Findings Are False" (Ioannidis 2005). Based on systematic reviews and simulations, the author concluded that "for most study designs and settings, it is more likely for a research claim to be false than true." Was this just an alarmist claim or is there indeed a problem with the validity of biomedical research? Despite some debate about the validity of Ioannidis' original analysis (e.g., Goodman and Greenland 2007), evidence has accumulated over the past 10 years that tends to favor the latter view. This is further supported by a recent commentary in *Nature* (Macleod 2011) that underscores concerns that experimental design and conduct need to improve in laboratory animal research.

Jeremy D. Bailoo, PhD, and Thomas S. Reichlin, PhD, are postdoctoral fellows and Hanno Würbel, PhD, is professor and head of the Division of Animal Welfare at the Veterinary Public Health Institute of the University of Bern, Switzerland.

Address correspondence and reprint requests to Hanno Würbel, Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Länggassstrasse 120, 3012 Bern, Switzerland or email hanno.wuerbel@vetsuisse.unibe.ch.

Poor Reproducibility and Translational Failure

The use of animals for research is a privilege granted to scientists with the explicit understanding that this use provides significant new knowledge without causing unnecessary harm. However, poor reproducibility of results from animal experiments across many research areas (c.f., Richter et al. 2009) and widespread failure to translate preclinical animal research to clinical trials (i.e., translational failure; e.g., Kola and Landis 2004; Howells et al. 2014; van der Worp et al. 2010) suggest that these expectations are not met. For example, of more than 500 neuroprotective interventions that were effective in animal models of ischemic stroke, none was found to be effective in humans (O'Collins et al. 2006). A 10-year review (1991–2000) of drug development revealed that the main causes of such attrition at the clinical trials stage are lack of efficacy and safety, which together account for 60% of the overall attrition rate (Kola and Landis 2004). These authors therefore concluded that animal studies which better predict the efficacy and safety of drugs in clinical trials are needed to reduce translational failure.

The Study of the Scientific Validity of Laboratory Animal Research

The empirical study of the scientific validity of laboratory animal research is an emerging field (Macleod 2011), and several lines of evidence highlight both current and potential problems. For example, translational failure in drug development could indicate that the construct validity of animal models is poor (Box 1). Construct validity refers to the degree to which a test measures what it claims to be measuring (Cronbach and Meehl 1955), and there is increasing concern that the construct validity of many animal models for human diseases is indeed questionable (e.g., Editor 2011; Nestler and Hyman 2010). However, construct validity depends on the specific disease that is modeled, and there is no simple method for assessing construct validity. Furthermore, improvements in animal models usually go hand in hand with

ILAR Journal, Volume 55, Number 3, doi: 10.1093/ilar/ilu037

Box 1. Glossary of Key Terms

Bias: Systematic deviation from the true value of the estimated treatment effect caused by failures in the design, conduct, or analysis of an experiment.

- **Attrition bias:** The unequal distribution of dropouts or nonresponders between treatment groups. This can lead to a systematic difference between treatment groups and may lead to an incorrect ascription of a causal relation between the treatment and the dependent variable.
- **Detection bias:** Systematic differences between treatment groups in how outcomes are assessed. This can be reduced or avoided by blinding or masking.
- **Performance bias:** Systematic differences in animal care and handling between treatment groups. This can be reduced or avoided by blinding or masking.
- **Selection bias:** The biased allocation of subjects to treatment groups. Biased allocation can lead to systematic differences in the baseline characteristics between groups. This can be avoided by randomized allocation and allocation concealment.

Blinding/masking: The maintenance of the persons' (who perform the experiment, collect data, and assess outcome, etc.) unawareness of the treatment allocation.

Types of error

- **False negative (β):** The failure to reject the null hypothesis when it is false. This is often due to small sample sizes (underpowered study designs).
- **False positive (α):** The rejection of the null hypothesis when it is true. This is often due to some form of bias.

Randomization

- **Simple:** Randomized allocation of subjects to the different treatment groups based on a single sequence of random assignments. This may lead to imbalanced groups and group sizes when the number of subjects is small.
- **Stratified:** Allocation of subjects to blocks of subjects sharing similar baseline characteristics (e.g., sex, age, body size) followed by randomized allocation of the subjects of each block to the different treatment groups. This is intended to counterbalance potential covariates across treatment groups.

Reproducibility: The ability of a result to be replicated by an independent experiment in the same or a different laboratory.

Validity

- **Construct validity:** The degree to which inferences are warranted from the sampling properties of an experiment (e.g., units, settings, treatments and outcomes) to the entities these samples are intended to represent.
- **External validity:** The extent to which the results of an animal experiment provide a correct basis for generalizations to other populations of animals (including humans) and/or other environmental conditions.
- **Internal validity:** The extent to which the design, conduct, and analysis of the experiment eliminate the possibility of bias so that the inference of a causal relationship between an experimental treatment and variation in an outcome measure is warranted.

Definitions adapted from van der Worp et al. (2010) and from the Cochrane Collaboration.

advances in research on the construct that is being modeled. Therefore, improving the construct validity of animal models depends on advances in research rather than adherence to methods or policies.

Another aspect related to construct validity concerns the health and well-being of the animals used for research. Growing evidence indicates that current standard practices of housing and care in laboratory animals are associated with abnormal brain and behavioral development and other signs of poor welfare, which may also compromise the scientific validity of research findings (Garner 2005; Knight 2001; Martin et al. 2010; Würbel 2001). Whether animal welfare matters in terms of the scientific validity of a research finding, however, depends on the area of research and on the specific research question.

Although highly relevant, construct validity and animal welfare will therefore not be further discussed in this article. Instead, we will focus our discussion on two fundamental aspects of scientific validity, both of which are relevant across all fields of laboratory animal research and are determined by experimental design and conduct: internal and external validity.

Internal and External Validity of Laboratory Animal Research

Internal validity refers to the extent to which a causal relation between an experimental treatment and variation in an outcome measure is warranted (Box 1). It critically depends on the extent to which experimental design and conduct minimize systematic error (also called bias). Already some 15 to 20 years ago, reports were published indicating that fundamental aspects of proper scientific conduct were often ignored, thereby compromising the internal validity of research findings (Festing and Altman 2002; McCance 1995). Several recent studies suggest that not much has changed to date. For example, a systematic review of animal experiments conducted in publicly funded research establishments in the United Kingdom and United States revealed that only a few authors reported using randomization (13%) or blinding (14%) to avoid bias in animal selection and outcome assessment (Kilkenny et al. 2009). Others found that only 3% of all studies reported an *a priori* sample size calculation (Sena et al. 2007) and in even fewer cases was a primary outcome variable defined (Macleod 2011). Similar results were obtained from various reviews of preclinical neurological research (Frantzias et al. 2011; van der Worp et al. 2010; Vesterinen et al. 2010), indicating that systematic bias may be widespread in laboratory animal research.

In clinical research, similar problems became apparent several years earlier, resulting in the CONSORT statement intended to improve the reporting of randomized clinical trials (Begg 1996; Moher et al. 2001; Schulz et al. 2010). Based on the CONSORT statement and with the aim to improve the reporting of animal studies, Kilkenny et al. (2010) recently developed the Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines, a 20-item checklist of information to be reported in publications of animal research. To date, these guidelines have been endorsed by over 430 journals, funders, universities, and learned societies (www.NC3Rs.org.uk) in the hope that such guidelines will not only improve the quality of scientific reporting but also the internal validity of the research.

In contrast to internal validity, external validity extends beyond the specific experimental setting and refers to the generalizability of research findings, i.e., how applicable they are to other environmental conditions, experimenters, study populations, and even to other strains or species of animals (including humans; Lehner 1996; Box 1). Poor external validity may thus contribute to both poor reproducibility of a research finding (e.g., when the same study replicated in a different laboratory by a different experimenter produces different results) and translational failure (e.g., when a treatment shown to be efficacious in an animal model is not efficacious in a clinical trial in humans).

Importantly, some of the strategies employed to increase internal validity may at the same time decrease external validity. For example, common strategies of standardizing experiments by using homogenous study populations to maximize test sensitivity inevitably compromise the external validity of the research findings, resulting in poor reproducibility (Richter et al. 2009, 2010, 2011; van der Worp et al. 2010; Würbel 2000, 2002; Würbel and Garner 2007).

Scope for Refinement of Laboratory Animal Research

Taken together, there seems to be considerable scope for refinement of experimental design and conduct to improve both the internal and external validity of laboratory animal research. In the following sections, we will explore this in more detail and propose potential ways of refinement as well as promising areas of future research.

Internal Validity – Refinement of Experimental Conduct to Avoid Systematic Biases

Although 235 different types of bias in biomedical research have been characterized (Chavalarias and Ioannidis 2010), van der Worp et al. (2010) consider four types of bias to be particularly relevant with respect to the internal validity of laboratory animal research: selection bias, attrition bias, performance bias, and detection bias.

Selection bias refers to the biased allocation of animals to treatment groups and can be avoided by randomization (Box 1). Because selection bias may occur either consciously or subconsciously, methods based on active decisions by the experimenter (e.g., picking animals "at random" from their cages) are not considered true randomization. Tossing coins or throwing dice provide simple ways of randomization but for some purposes random number generators (e.g., www.random.org) may be preferable. Even the use of allegedly "homogeneous" study populations (such as same-sex, same-age inbred mice raised under identical housing conditions) does not preclude the need for randomization, because individual differences still prevail. This is best illustrated by studies with inbred mice showing that variation within strains is often significantly greater than between strains (Wahlsten 2010).

In many cases, it is possible to use stratified randomization instead of simple randomization. In stratified randomization, the study population is divided into discrete subpopulations based on systematic differences in factors that are likely to affect the outcome measures, such as sex, age, littermates, disease severity, treatment dose, etc. The animals of each subpopulation are then separately allocated at random to the different treatment groups. Through this, the factor levels defining the different subpopulations are counterbalanced among all treatment groups. The use of statistical methods designed to analyze such factorial designs results in the removal of the variation between the strata from the error term, thereby increasing the precision and statistical power of the experiment (Altman and Bland 1999).

Selection bias may also occur when the criteria for inclusion or exclusion of animals are poorly defined. Complications that require exclusion of animals are an inherent risk in animal studies, especially with animal models involving invasive surgical procedures (e.g., Jüni et al. 2001) and in models of stroke (Crossley et al. 2008). For ethical reasons, humane endpoints need to be defined *a priori*, and animals that reach humane endpoints may be lost from the subsequent analysis. However, it may also be justifiable to exclude animals for scientific reasons if complications occur that are unrelated to the experimental treatment and render the outcome measures meaningless. To avoid bias, however, all criteria for inclusion and exclusion of animals need to be predefined, and the person deciding on inclusion or exclusion needs to be unaware of the treatment allocation (van der Worp et al. 2010). If these criteria are not well specified, one risks the induction of attrition bias, the unequal distribution of dropouts among treatment groups.

Performance bias may occur whenever there is a systematic difference in the interaction with the animals (e.g., animal care, experimental procedures) between the treatment groups, apart from the treatment under investigation (Jüni et al. 2001; Box 1). For example, differences in the quality of experimenter handling exhibited to stressed vs. nonstressed mice may occur due to higher fearfulness and stress reactivity in the stressed mice (Hurst and West 2010). In contrast, detection bias occurs when the outcome is measured differently in animals of different treatment groups. Again, both performance bias and detection bias may occur either consciously or subconsciously, and the best way to avoid these biases is blinding (also known as masking).

Blinding is considered complete when the investigator and everyone else involved in the experiment (animal care personnel, laboratory technicians, outcome assessors, etc.) are unaware of the animals' allocation to treatments. In contrast to randomization, blinding is not always possible, for example, when scoring behavior among treatment groups that differ visibly (e.g., strains of mice differing in coat color). Thus, it is important that authors explicitly report the blinding status of all people whose involvement may affect the outcome of the study (Kilkenny et al. 2010a; Moher et al. 2010).

Other relevant sources of bias include sample sizes that are either too small or too large, a poor definition of the primary (and secondary) outcome variable(s), and the use of inappropriate statistical analyses, all of which may result in poor statistical conclusion validity (Cozby and Bates 2011). Whenever possible, a formal sample size calculation (and power analysis) should be performed that specifies the minimal effect size considered to be relevant (e.g., Cohen's d or f), the desired statistical power (1–β), and the level of statistical significance (α). Some have argued that such calculations are only applicable to "confirmatory research" but not to "exploratory research" since "effect sizes may be unknown" and "research in the exploratory mode will often test many different strategies in parallel, and this is only feasible if small sample sizes are used" (Kimmelman et al. 2014). However, neither unknown effect sizes nor the exploratory nature of research should be taken as excuses for violating fundamental principles of good scientific practice. Tools such as NCSS PASS, G*Power, and the resource equation method (Mead 1990) (to name just a few) facilitate sample size calculations. This is even possible when knowledge about the sample distribution is incomplete because usually a minimally relevant effective size can be specified *a priori*. Furthermore, testing many different hypotheses in parallel using small sample sizes will inevitably produce spurious results that undermine the reliability of the research (Button et al. 2013).

Both overpowered and underpowered studies are unethical, albeit for different reasons. Overpowered studies use more animals than needed to detect a significant effect of a given size. This is relatively rare, however, because it violates one of the 3R principles (reduction), and ethics committees are trained to spot reduction potential. From a scientific perspective, large sample sizes are not a problem as such, as long as a minimal effect size is defined. However, any two treatments will be significantly different if the measurement precision and sample size are large enough, and so overpowered designs may lead to bias when biologically irrelevant effect sizes are considered relevant because of their statistical significance. Underpowered studies are much more prevalent, even though they are much more problematic from both ethical and scientific points of view (c.f., Button et al. 2013). Underpowered studies are unable to detect biologically relevant effect sizes, and as a result, the animals are essentially wasted for inconclusive research. On the other hand, there are obvious economic incentives to keep sample sizes small. In addition, it appears that the well-intended yet one-sided focus of ethics committees on reduction may further promote underpowered study designs (Demétrio et al. 2013). In the human clinical trial literature, the ethical and scientific costs of underpowered study designs have long been recognized (Halpern 2002); it is crucial that formal power calculations become standard practice in animal research so that scientific gain is maximized while animal use is minimized (Button et al. 2013; Kilkenny et al. 2010b; Macleod 2011).

Recent evidence from preclinical neurological research indicates that there are also too many statistically significant (i.e., "positive") results in the literature (Ioannidis and Trikalinos 2007; Tsilidis et al. 2013). These authors concluded that selective analysis and selective outcome reporting are the most likely causes. Selective analysis occurs when several statistical analyses are performed but only the one with the "best" (i.e., most significant) result is presented (Ioannidis 2008; Tsilidis et al. 2013). Similarly, selective outcome reporting occurs when many outcome variables are analyzed but only the variables that are significantly affected by the treatment are reported (Tsilidis et al. 2013). While the possible merits of selective reporting are still debated (e.g., de Winter and Happee 2013; van Assen et al. 2014), we maintain that to avoid these potential biases, the primary (and secondary) outcome variable(s) as well as the statistical approach(es) to testing for treatment effects need to be

specified before the onset of the study. Ultimately, the best way to achieve this would be the prospective registration of all animal studies (see below).

Finally, as the use of the scientific method requires reproducibility and falsifiability, the sharing of collected data (i.e., public data archiving) and validation of published analytical methods should become more common (Molloy 2011). Although this topic is not without issue or debate (e.g., Alsheikh-Ali et al. 2011; Editor 2014; Nelson 2009; Roche et al. 2014), the transparency of collected data can only improve the quality of published scientific results.

Do Reporting Guidelines Help?

The common approach to reducing poor experimental conduct has been the implementation of reporting guidelines. This started with the CONSORT statement to improve the reporting of human clinical trials about 20 years ago (Begg 1996; Moher et al. 2001; Schulz et al. 2010) and was recently extended to animal research by the ARRIVE guidelines (Kilkenny et al. 2010b). Similar reporting guidelines are available for other areas of research, such as STROBE for epidemiology (von Elm et al. 2007), PRISMA for systematic reviews and meta-analyses (Moher 2009), and several others listed by the EQUATOR Network (www.equator-network.org). More recently, it has been proposed that animal experiments should be preregistered (Chambers 2013), similarly to clinical trials which according to the Declaration of Helsinki (WMA 2013) must be registered in a publicly accessible database (e.g., www.ClinicalTrials.gov) before recruitment of the first subject. Preregistration should help to avoid "inappropriate research practices, including inadequate statistical power, selective reporting of results, undisclosed analytic flexibility, and publication bias" (Chambers 2013). All of these initiatives reflect the pervasive nature of bias in biomedical research.

So, do reporting guidelines improve experimental conduct? Although there is only indirect evidence, there is good reason to believe that they do indeed. For example, systematic reviews and meta-analyses in preclinical research on stroke, multiple sclerosis, and Parkinson's disease indicate that poor reporting of study quality attributes (e.g., randomization, blinding, sample size calculation, etc.) correlates with overstated treatment effects (Rooke et al. 2011; Sena et al. 2007; Vesterinen et al. 2010). It is therefore plausible that better reporting correlates with better quality of study conduct. Although, theoretically, the reporting of accurate study quality may be faked, such outright fraud is hopefully uncommon. It is more likely that the advocacy of reporting guidelines will raise awareness of the importance of rigorous experimental conduct (Landis et al. 2012). Nevertheless, a recent analysis of papers published in the *PLoS* and *Nature* journals after the endorsement of the ARRIVE guidelines found as yet very little improvement in reporting standards, indicating that authors are still ignoring, and referees and editors are not enforcing, these guidelines (Baker et al. 2014).

External Validity – Refinement of Experimental Design to Avoid Spurious Results

Reproducibility is a cornerstone of the scientific method, and poor reproducibility threatens the credibility of the entire field of animal research (Johnson 2013; Richter et al. 2009). Although better internal validity will also improve the reproducibility of results, reproducibility of a result is primarily a function of external validity (Richter et al. 2009; Würbel 2000).

By definition, external validity refers to the applicability of results to other environmental conditions, experimenters, study populations, and even to other strains or species of animals (including humans; Lehner 1996; Box 1). External validity therefore defines how generalizable results are. This also includes reproducibility, which is defined as the ability of a result to be replicated by an independent experiment either in the same or in a different laboratory (Box 1). However, the relationship between external validity and reproducibility is not so straightforward. External validity (i.e., the range of conditions to which a result can be generalized) is an inherent feature of a result; some results are more externally valid than others. For example, pre-pulse inhibition (PPI) of the startle reflex to acoustic stimuli is highly conserved across many species, including mice and humans, and is fairly robust against variation in environmental conditions (Geyer et al. 2002). Thus, PPI has very high external validity. Because of this, PPI is also highly reproducible across different laboratories despite considerable variation in conditions among laboratories. In contrast, the locomotor activity of mice on an elevated zero-maze or plus-maze has very little external validity, as it is highly sensitive to test conditions (e.g., handling; Hurst and West 2010), and differences between strains of mice are highly inconsistent despite considerable efforts to equate conditions across laboratories (e.g., Crabbe et al. 1999; Richter et al. 2011). Therefore, experiments should be designed in ways that permit for estimation of the external validity of the results. This can only be achieved if relevant features of the study design, such as animal characteristics and environmental conditions, are varied systematically (Würbel 2000, 2002).

Interestingly, this is contrary to conventional wisdom in laboratory animal science. The gold standard of experimental design adopted from the pure sciences (mathematics, physics, chemistry) is to hold constant all factors except for the independent variable(s) under investigation. This has become a central dogma in laboratory animal science that is referred to as standardization. Thus, laboratory animal science textbooks advise researchers to standardize their experiments by using genetically uniform animals, selecting these for maximal phenotypic uniformity (e.g., same age, same weight, etc.), and keeping all environmental and procedural factors constant (Beynen, Festing, et al. 2001; Beynen, Gärtner, et al. 2001). Such homogenization of study populations may compromise both the external validity and reproducibility of the results, an effect that has been referred to as the

standardization fallacy (Würbel 2000, 2002). The same fallacy was highlighted 80 years ago by the eminent Ronald A. Fisher (1935, p. 102): "The exact standardisation of experimental conditions, which is often thoughtlessly advocated as a panacea, always carries with it the real disadvantage that a highly standardised experiment supplies direct information only with respect to the narrow range of conditions achieved by standardisation. Standardisation, therefore, weakens rather than strengthens our ground for inferring a like result, when, as is invariably the case in practice, these conditions are somewhat varied."

Indeed, despite rigorous standardization of the experimental conditions across laboratories, several multi-laboratory studies revealed large proportions of results that were idiosyncratic to one laboratory (Crabbe et al. 1999; Richter et al. 2011; Wolfer et al. 2004). The reason for this may be that many environmental factors (e.g., staff, noise, etc.) cannot be equalized between laboratories, so that different laboratories inevitably standardize to different local environments (Richter et al. 2009; Würbel and Garner 2007). Therefore, standardization may actually be a cause of, rather than a cure for, poor reproducibility (Richter et al. 2009). Thus, not surprisingly, van der Worp et al. (2010) listed homogenous study populations as a main source of poor external validity in preclinical animal research, which to some extent may also contribute to translational failure.

Some scientists have argued that we simply need to report more parameters that may potentially affect outcome measures (e.g., Arndt and Surjo 2001; Philip et al. 2010; Surjo and Arndt 2001). In this case, however, reporting guidelines will not help, and the attempt to promote extensive lists of methodological detail to facilitate interpretation of conflicting findings has been referred to as the listing fallacy (Würbel 2002). If anything, such lists may induce interpretation bias by attracting attention to differences in the listed parameters, although there may be many more parameters that were not considered, were considered to be irrelevant or too difficult to assess, or simply could not be listed. As long as a particular parameter has not been varied systematically within a given experiment, it is no more likely to explain conflicting findings than any other parameter, listed or unlisted, that differed between the respective experiments (Würbel 2002).

Statistical and Experimental Solutions

Among studies investigating behavioral differences between different inbred and mutant strains of mice (behavioral phenotyping), current estimates of the proportion of irreproducible results (false discovery proportion, FDP) from multi-laboratory studies range between 30 and 60% (Benjamini et al. 2014; Kafkafi et al. 2005, 2014). It is likely that similar FDPs apply to other areas of research.

Various solutions have been proposed to reduce the risk of obtaining such spurious results. For example, Johnson (2013) suggested lowering the critical *P* value of statistical significance from 0.05 to 0.005 or even 0.001 to match conventional evidence thresholds used in Bayesian testing. Assuming that approximately one-half of the hypotheses tested by scientists are true, Johnson (2013) estimated that between 17% and 25% of marginally significant scientific findings are false positives. However, to avoid lowering the proportion of false positives and increasing the proportion of false negatives, sample sizes would have to be increased by about 50% to 100% to achieve similar statistical power (Johnson 2013). Moreover, a general decrease of critical *P* values does not take into account that both external validity and reproducibility depend on the nature of the measured effect.

Kafkafi and colleagues (2005, 2014) have therefore proposed to raise the benchmark for significant results in a more specific way. According to their random laboratory model, laboratories should be considered as a sample, representing the population of all potential laboratories, and the interaction noise (the treatment x laboratory variance) should be added as a random factor to the individual animal noise (the within-laboratory variance). Similarly to the suggestion of lowering *P* values (Johnson 2013), this inflation of within-laboratory variance would generate a larger yardstick for the significance of treatment effects (Benjamini et al. 2014; Kafkafi et al. 2005, 2014), albeit in a more specific way. Using data from several multi-laboratory studies, the authors showed that this method may reduce the FDP considerably without losing too much statistical power. The difficulty with this approach is that such specificity will be achieved only if the treatments and measures are first tested across several laboratories to obtain accurate estimates of between-laboratory variance. This approach may thus not be applicable to animal experiments in general but may be useful for standard preclinical tests of efficacy and toxicity in drug development, as well as for specific large scale projects, such as the International Mouse Phenotyping Consortium (Brown and Moore 2012a, 2012b; Mallon et al. 2012) which aims to determine the phenotypes of thousands of mutant lines with a battery of standard tests (Benjamini et al. 2014; Kafkafi et al. 2005, 2014).

Besides these statistical approaches, others have proposed mimicking between-laboratory variability experimentally. These proposals range from conducting an independent replicate study to conducting real multi-laboratory studies. For example, the Reproducibility Initiative has established a service to facilitate independent replicate studies (http://validation.scienceexchange.com/), while the MultiPART consortium aims to develop a platform for international multicenter preclinical stroke trials based on randomized clinical trial design (www.dcn.ed.ac.uk/multipart/).

In addition to such true replications, there are several other ways in which studies may be designed to provide an estimate of the external validity and reproducibility of results. For example, Richter and colleagues (2010, 2011) proposed the heterogenization of study populations (rather than homogenization through standardization) by systematically varying a few selected factors. In principle, any aspect of the animals (e.g., genotype, sex, age, body condition, etc.) and their environment (e.g., housing conditions,

experimental protocol) may be used for such heterogenization. By varying two environmental factors using a 2 × 2 factorial design, Richter and colleagues (2010) successfully mimicked variation between independent replicates conducted within their own laboratory (see also Jonker et al. 2013; Wolfinger 2013; Würbel et al. 2013). However, a similar simple form of heterogenization did not account for between-laboratory variation in a true multi-laboratory study (Richter et al. 2011). Further research is therefore needed to develop heterogenization protocols that mimic between-laboratory variability more effectively.

In the meantime, simple precautions may be taken as proposed by Paylor (2009), for example, by splitting experiments into small batches of animals that are tested some time apart instead of testing them in one large batch, by using multiple experimenters for testing and data collection instead of using only one, or by spreading test sessions across time of day instead of testing all animals at the same time of day. Assessing the effects of batch, experimenter, or time of day, respectively, will reveal whether such minor variations of conditions affect results and will therefore indicate whether reproducibility across larger variations of conditions (such as between laboratories) may be at stake.

Conclusions

Reproducibility and falsifiability are cornerstones of the scientific method, and it is because of these principles that science is often viewed as self-correcting, at least in the long term. The common consensus is that failures in reproducibility of animal research are not a consequence of scientific misconduct (e.g., Collins and Tabak 2014). However, negligence in experimental design, conduct, and publication (whether conscious or not) continue to plague animal research and, despite numerous initiatives to curb these effects, they continue to persist (Baker et al. 2014).

Facing these problems and underlying causes, as discussed here, is hopefully a step towards effective refinement of experimental design and conduct. The ARRIVE guidelines provide a useful tool for improving the internal validity of animal research, and several strategies have been put forward for improving external validity as well. Nevertheless, it seems that greater pressure must be placed on researchers, reviewers, and journal editors to not only endorse such methods of refinement but to rigorously enforce them. Otherwise, the credibility and ethical justification of animal research may be permanently undermined.

Acknowledgments

The authors of this paper were funded by the ERC Advanced Grant "REFINE" (H.W. and J.D.B.), the FP7 Coordination and Support Action "Multi-PART" (H.W. and J.D.B.), and a research grant by the Swiss Federal Food Safety and Veterinary Office (H.W. and T.S.R.).

References

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. 2011. Public availability of published research data in high-impact journals. PLoS One 6:e24357.

Altman DG, Bland JM. 1999. How to randomise. BMJ 319:703–704.

Arndt SS, Surjo D. 2001. Methods for the behavioural phenotyping of mouse mutants. How to keep the overview. Behav Brain Res 125:39–42.

Baker D, Lidster K, Sottomayor A, Amor S. 2014. Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. PLoS Biol 12:e1001756.

Begg C. 1996. Improving the quality of reporting of randomized controlled trials. JAMA 276:637–639.

Benjamini Y, Lahav T, Kafkafi N. 2014. Estimating replicability of behavioral phenotyping results in a single laboratory. In: Measuring Behavior 2014: The replicability of measuring behavior.

Beynen AC, Festing MFW, van Montfort MAJ. 2001. Design of animal experiments. In: Van Zutphen LFM, Baumans V, Beynen AC, eds. Principles of Laboratory Animal Science. Revised. Amsterdam: Elsevier. p 219–249.

Beynen AC, Gärtner K, van Zutphen LFM. 2001. Standardization of Animal Experimentation. In: Van Zutphen LFM, Baumans V, Beynen AC, eds. Principles of Laboratory Animal Science. Revised. Amsterdam: Elsevier. p 103–110.

Brown SDM, Moore MW. 2012a. The International Mouse Phenotyping Consortium: Past and future perspectives on mouse phenotyping. Mamm Genome 23:632–640.

Brown SDM, Moore MW. 2012b. Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. Dis Model Mech 5:289–292.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Chambers CD. 2013. Registered reports: A new publishing initiative at Cortex. Cortex 49:609–610.

Chavalarias D, Ioannidis JPA. 2010. Science mapping analysis characterizes 235 biases in biomedical research. J Clin Epidemiol 63:1205–1215.

Collins FS, Tabak LA. 2014. Policy: NIH plans to enhance reproducibility. Nature 505:612–613.

Cozby P, Bates S. 2011. Methods in Behavioral Research. 11th ed. McGraw-Hill Education.

Crabbe JC, Wahlsten DL, Dudek BC. 1999. Genetics of Mouse behavior: interactions with laboratory environment. Science 284:1670–1672.

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. Psychol Bull 52:281–302.

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PMW, Macleod M, Dirnagl U. 2008. Empirical evidence of bias in the design of experimental stroke studies: A metaepidemiologic approach. Stroke 39:929–934.

Demétrio CGB, Menten JFM, Leandro RA, Brien C. 2013. Experimental power considerations-justifying replication for animal care and use committees. Poult Sci 92:2490–2497.

De Winter J, Happee R. 2013. Why selective publication of statistically significant results can be effective. PLoS One 8:e66463.

Editor. 2011. Building a better mouse test. Nat Methods 8:697.

Editor. 2014. Share alike. Nature 507:140.

Festing MFW, Altman DG. 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J 43:244–258.

Fisher RA. 1935. The design of experiments. Oliver & Boyd.

Frantzias J, Sena ES, Macleod MR, Al-Shahi Salman R. 2011. Treatment of intracerebral hemorrhage in animal models: Meta-analysis. Ann Neurol 69:389–399.

Garner JP. 2005. Stereotypies and Other abnormal repetitive behaviors: Potential impact on validity, reliability, and replicability of scientific outcomes. ILAR J 46:106–117.

Geyer MA, McIlwain KL, Paylor R. 2002. Mouse genetic models for prepulse inhibition: An early review. Mol Psychiatry 7:1039–1053.

Goodman S, Greenland S. 2007. Why most published research findings are false: Problems in the analysis. PLoS Med 4:e168.
Halpern SD. 2002. The continuing unethical conduct of underpowered clinical trials. JAMA 288:358–362.
Howells DW, Sena ES, Macleod MR. 2014. Bringing rigour to translational medicine. Nat Rev Neurol 10:37–43.
Hurst JL, West RS. 2010. Taming anxiety in laboratory mice. Nat Methods 7:825–826.
Ioannidis JPA. 2005. Why most published research findings are false. PLoS Med 2:e124.
Ioannidis JPA. 2008. Why most discovered true associations are inflated. Epidemiology 19:640–648.
Ioannidis JPA, Trikalinos TA. 2007. The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. CMAJ 176:1091–1096.
Johnson VE. 2013. Revised standards for statistical evidence. Proc Natl Acad Sci U S A 110:19313–19317.
Jonker RM, Guenther A, Engqvist L, Schmoll T. 2013. Does systematic variation improve the reproducibility of animal experiments? Nat Methods 10:373.
Jüni P, Altman DG, Egger M. 2001. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 323:42–46.
Kafkafi N, Benjamini Y, Sakov A, Elmer GI, Golani I. 2005. Genotype-environment interactions in mouse behavior: A way out of the problem. Proc Natl Acad Sci U S A 102:4619–4624.
Kafkafi N, Lahav T, Benjamini Y. 2014. What's always wrong with my mouse? In: Measuring Behavior 2014: The replicability of Measuring Behavior.
Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010a. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 8:e1000412.
Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010b. Animal research: reporting in vivo experiments: The ARRIVE guidelines. J Gene Med 12:561–563.
Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One 4:e7824.
Kimmelman J, Mogil JS, Dirnagl U. 2014. Distinguishing between exploratory and confirmatory preclinical research will improve translation. Jones DR, editor. PLoS Biol 12:e1001863.
Knight J. 2001. Animal data jeopardized by life behind bars. Nature 412:669.
Kola I, Landis J. 2004. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3:711–715.
Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT 3rd, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490:187–191.
Lehner PN. 1996. Handbook of Ethological Methods. Cambridge University Press.
Macleod M. 2011. Why animal research needs to improve. Nature 477:511.
Mallon A-M, Iyer V, Melvin D, Morgan H, Parkinson H, Brown SDM, Flicek P, Skarnes WC. 2012. Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. Mamm Genome 23:641–652.
Martin B, Ji S, Maudsley S, Mattson MP. 2010. "Control" laboratory rodents are metabolically morbid: why it matters. Proc Natl Acad Sci U S A 107:6127–6133.
McCance I. 1995. Assessment of statistical procedures used in papers in the Australian Veterinary Journal. Aust Vet J 72:322–329.
Mead R. 1990. The design of experiments: Statistical principles for practical applications. Cambridge University Press.
Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. 2010. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. BMJ 340:c869.
Moher D, Schulz KF, Altman DG. 2001. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357:1191–1194.
Moher D. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Ann Intern Med 151:264.
Molloy JC. 2011. The Open Knowledge Foundation: open data means better science. PLoS Biol 9:e1001195.
Nelson B. 2009. Data sharing: Empty archives. Nature 461:160–163.
Nestler EJ, Hyman SE. 2010. Animal models of neuropsychiatric disorders. Nat Neurosci 3:1161–1169.
O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 2006. 1,026 experimental treatments in acute stroke. Ann Neurol 59:467–477.
Paylor R. 2009. Questioning standardization in science. Nat Methods 6:253–254.
Philip VM, Duvvuru S, Gomero B, Ansah TA, Blaha CD, Cook MN, Hamre KM, Lariviere WR, Matthews DB, Mittleman G, Goldowitz D, Chesler EJ. 2010. High-throughput behavioral phenotyping in the expanded panel of BXD recombinant inbred strains. Genes Brain Behav 9:129–159.
Richter SH, Garner JP, Auer C, Kunert J, Würbel H. 2010. Systematic variation improves reproducibility of animal experiments. Nat Methods 7:167–168.
Richter SH, Garner JP, Würbel H. 2009. Environmental standardization: Cure or cause of poor reproducibility in animal experiments? Nat Methods 6:257–261.
Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, Touma C, Schindler B, Chourbaji S, Brandwein C, Gass P, van Stipdonk N, van der Harst J, Spruijt B, Võikar V, Wolfer DP, Würbel H. 2011. Effect of population heterogenization on the reproducibility of mouse behavior: A multi-laboratory study. PLoS One 6:e16461.
Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, Kruuk LEB. 2014. Troubleshooting public data archiving: Suggestions to increase participation. PLoS Biol 12: e1001779.
Rooke EDM, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. 2011. Dopamine agonists in animal models of Parkinson's disease: A systematic review and meta-analysis. Parkinsonism Relat Disord 17:313–320.
Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. BMC Med 8:18.
Sena E, van der Worp HB, Howells D, Macleod M. 2007. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci 30:433–439.
Surjo D, Arndt SS. 2001. The Mutant Mouse Behaviour network, a medium to present and discuss methods for the behavioural phenotyping. Physiol Behav 73:691–694.
Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, Al-Shahi Salman R, Macleod MR, Ioannidis JPA. 2013. Evaluation of excess significance bias in animal studies of neurological diseases. PLoS Biol 11:e1001609.
Vesterinen HM, Sena ES, Ffrench-Constant C, Williams A, Chandran S, Macleod MR. 2010. Improving the translational hit of experimental treatments in multiple sclerosis. Mult Scler 16:1044–1055.
Wahlsten DL. 2010. Mouse Behavioral Testing: How to use Mice in Behavioral Neuroscience. First. Elsevier.
WMA. 2013. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. JAMA 310:2191–2194.
Wolfer DP, Litvin O, Morf S, Nitsch RM, Lipp H-P, Würbel H. 2004. Laboratory animal welfare: Cage enrichment and mouse behaviour. Nature 432:821–822.
Wolfinger RD. 2013. Reanalysis of Richter et al. (2010) on reproducibility. Nat Methods 10:373–374.

Van Assen MALM, van Aert RCM, Nuijten MB, Wicherts JM. 2014. Why publishing everything is more effective than selective publishing of statistically significant results. PLoS One 9:e84896.

Van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. 2010. Can animal models of disease reliably inform human studies? PLoS Med. 7:e1000245.

Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Prev Med (Baltim) 45:247–251.

Würbel H. 2000. Behaviour and the standardization fallacy. Nat Genet 26:263.

Würbel H. 2001. Ideal homes? Housing effects on rodent brain and behaviour. TRENDS Neurosci 24:207–211.

Würbel H. 2002. Behavioral phenotyping enhanced--Beyond (environmental) standardization. Genes Brain Behav 1:3–8.

Würbel H, Garner JP. 2007. Refinement of rodent research though environmental enrichment and systematic randomization. Available from: http://www.nc3rs.org.uk/downloaddoc.asp?id=506&page=395&skin=0

Würbel H, Richter SH, Garner JP. 2013. Reply to: "Reanalysis of Richter et al. (2010) on reproducibility". Nat Methods 10:374.