More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research

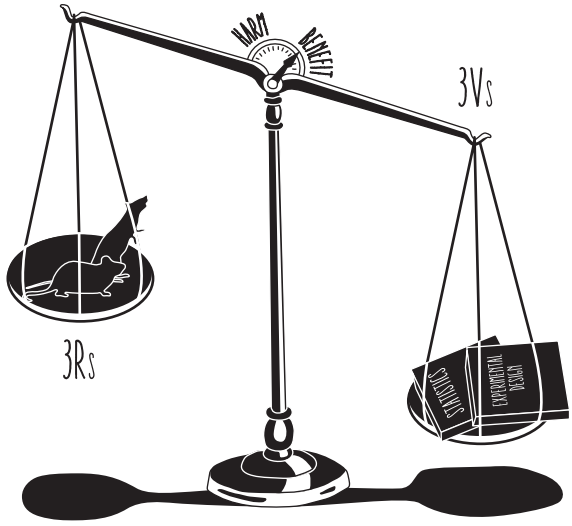
Hanno Würbel

The reproducibility crisis in biomedical research presents a new challenge for conducting harm-benefit analysis: how do we improve the validity of studies to maximize the likelihood of benefit?

Every year, 50–100 million vertebrates are used in experimental procedures worldwide. The use of animals for research is legally regulated on the explicit understanding that such use will provide significant new knowledge facilitating relevant benefits, and no unnecessary harm will be imposed on the animals[1]. Harm-benefit analysis (HBA) is the common tool for making ultimate decisions on whether study protocols meet these expectations. Therefore, HBA is a crucial part of project evaluation and explicitly required by the EU Directive 2010/63; it is also implied in the US Guide for the Care and Use of Laboratory Animals and emphasized in the Terrestrial Animal Health Code by the World Organization for Animal Health (OIE)[2].

HBA follows the legal principle of proportionality and involves three main questions, namely (1) whether the study is suitable for achieving a legitimate aim, (2) whether it is necessary, and (3) whether it is adequate. Question (3) refers to the actual HBA, which evaluates whether the expected benefits of a study outweigh the harms imposed on the animals. Questions (1) and (2) are instrumental prerequisites for the actual HBA; they are concerned with the scientific rationale underpinning the expected outcome of the study (suitability) and potential alternatives to the likely harms imposed on the animals (necessity).

Evaluation of potential alternatives essentially examines whether the 3Rs principle[3] has been exploited to minimize the harms imposed on the animals. Thus, for a study protocol to proceed to the final HBA, it must argue convincingly that the expected outcome cannot be achieved by using no or non-sentient animals (replace), by using fewer animals (reduce), or by using less harmful procedures (refine). In particular, refinements such as enriched housing, habituation to procedures, non-invasive techniques, and anesthetics and analgesics can shift weights in HBA of animal experiments by minimizing the harms imposed on the animals.



Kim Caesar/Springer Nature

FIGURE 1 | Refined procedure for harm-benefit-analysis (HBA) in animal research. Whereas 3Rs methods minimize the weight of harms to the animals on the HBA balance, methods to improve the scientific validity of the research (3Vs) maximize the value of study outcomes, thereby facilitating the expected benefits.

Bumping up the benefits

But what about the benefit side of the equation? Unless a study produces results that are scientifically valid and reproducible, the animals may be wasted for inconclusive research, no matter how little harm is inflicted on them[1]. Whereas 3R efforts to minimize harms to the animals are carefully scrutinized by ethical review committees, the scientific validity and reproducibility of study outcomes are generally taken for granted[4]. Such confidence may not be warranted as highlighted by the ongoing "reproducibility crisis" in biomedical research.

Over the past decade, evidence has accumulated indicating that scientific validity and reproducibility are alarmingly poor throughout biomedical research[1,5]. Based on systematic reviews and simulations, Ioannidis concluded that "for most study designs and settings, it is more likely for a research claim to be false than true"[6]. This is supported by evidence for risks of bias throughout *in vivo* research[4,7,8], spectacular cases of irreproducibility[9,10], and translational failure on a large scale[11,12].

Systematic error (bias), poor reproducibility, and translational failure can be caused by flaws at all levels of research, including

Division of Animal Welfare, Veterinary Public Health Institute, Vetsuisse Faculty, University of Bern, Länggassstrasse 120, 3012 Bern, Switzerland. Correspondence should be addressed to H.W. (hanno.wuerbel@vetsuisse.unibe.ch)

design, conduct, analysis, and reporting of experiments. For example, studies may use poorly validated animal models or outcome variables[13]; they may be based on samples that are too small[14] or idiosyncratic[15]; they may violate principles of good research practice (for example, randomization, blinded outcome assessment, a priori sample size calculation)[4,7,8] or use inappropriate statistics (for example, *p*-hacking)[16]; or they may report results selectively[17] or not at all (for example, publication bias)[18].

All of this can be detrimental to the scientific validity and reproducibility of results published in the primary scientific literature, thereby compromising the outcome of the research. In much the same way as the 3Rs principle serves to implement strategies that minimize harms to the animals, a more powerful principle may be needed to implement strategies that maximize scientific validity, thereby facilitating the benefits of animal experiments. The following analogy may illustrate this. When refinements for a harmful procedure are available (for example, post-surgical analgesia) but ignored in a study protocol, this represents a violation of the 3Rs principle, thereby causing unnecessary harms to the animals. Similarly, ignorance of measures against risks of bias (for example, randomization, blinded outcome assessment) can be regarded as violation of the principles of good research practice, thereby compromising the outcome of studies. However, similar to unavoidable harms, not all risks of bias are avoidable. For example, when assessing behavioral differences between mice of different coat color, blinded outcome assessment may be impossible. Although non-blinded outcome assessment represents a risk of bias that compromises the study outcome, it is not unethical. By contrast, when blinded outcome assessment is feasible but ignored without justification, it represents a case of irresponsible use of animals, which is unethical, and for example, in the EU is actually against the law.

There is some debate as to whether scientific validity should be weighed on the harm side or the benefit side of the equation[2], or whether it should be part of an independent third dimension "likelihood of benefit" as in "Bateson's cube"[19]. However, in their recent report on current concepts of HBA of animal experiments, the AALAS-FELASA Working Group concluded that "performing HBA in a systematic way and thereby defining and describing benefits is not common practice", but that "a well-designed experiment is a fundamental criterion for reliable information and for generating any benefit at all"[2].

TABLE 1 | Considerations for harm-benefit analysis

Suitable	**Assess validity (3Vs)**
	construct validity
	internal validity
	external validity
Necessary	**Assess harms (3Rs)**
	replace
	reduce
	refine
Adequate	**Harm-benefit analysis**
	weight of harms: consider 3Rs
	weight of benefits: consider 3Vs
	conduct harm-benefit analysis

The 3Vs of scientific validity

I therefore propose to extend HBA by adding a more systematic assessment of scientific validity and suggest including three key aspects of scientific validity, namely construct validity (cV), internal validity (iV), and external validity (eV), which for reasons of convenience I will hereafter refer to as the 3Vs. Thus, before the actual HBA, study protocols should not only be assessed for the 3Rs but also for the 3Vs (**Fig. 1** and **Table 1**). Assessment of construct validity should be based on evidence about the level of agreement between the animal model, test or outcome variable and the quality it is meant to measure[20]. In the case of outcome variables this may include evidence of convergent and discriminant validity; in the case of animal models for specific conditions (for example, diseases) in humans or other animals this may include evidence of the three main aspects of model validity: face, construct, and predictive validity[20,21]. Assessment of internal validity should be based on evidence for the scientific rationale (*e.g.* use of appropriate control groups) and for scientific rigor in terms of measures against risks of bias (for example, definition of primary and secondary outcome variables, sample size calculation, randomization, blinding, statistical analysis plan)[1,22]. Finally, assessment of external validity should be based on evidence for experimental design features that enhance, or facilitate inference about, the reproducibility and generalizability of the expected results[1]. This includes splitting experiments into multiple independent replicates (batches)[23], introducing systematic variation (heterogenization) of relevant variables (for example, species/strains of animals, housing conditions, tests, etc.)[15,24,25], or implementing multi-center study designs[26]. In this way, the 3Vs could offer welcome guiding principles for assessing and maximizing the scientific validity of study outcomes, thereby increasing the likelihood of achieving the expected benefit of animal experiments.

At present, ethical review does not include a systematic assessment of scientific validity in the course of HBA. For animal research in Switzerland we recently demonstrated that the authorities licensing animal experiments would actually lack important information to do so; the application form does not explicitly ask for it and, therefore, applicants do not provide it[4,8]. In light of the current "reproducibility crisis", I propose that a more systematic assessment of the 3Vs – similar to the assessment of the 3Rs – as part of HBA would provide a powerful tool to evaluate and enhance the scientific validity and reproducibility of *in vivo* research.

This seems particularly pertinent in terms of reproducibility and generalizability of research findings. The scope of animal experiments is often very narrow, most studies being conducted as small-scale single-laboratory studies. Due to the highly standardized conditions within laboratories, results of single-laboratory studies have often very little external validity[1,15,27]. Ironically, 3R efforts to minimize animal use (reduce) may inadvertently exacerbate this situation by promoting standardization as a means to reduce within-experiment variation in view of smaller sample sizes[28]. However, this can be counterproductive since standardization inevitably reduces external validity, and as a consequence reproducibility[27,29].

Using data from 50 independent studies on the effect of hypothermia on infarct volume in animal models of stroke, we recently conducted a simulation study to analyze reproducibility of single-laboratory studies compared to multi-laboratory studies. Treatment

effects of single-laboratory studies varied widely (between 0% and 100% reduction of infarct volume), and this variation was reduced considerably by multi-laboratory designs. Furthermore, whereas less than 50% of single-laboratory studies produced an accurate estimate of the "true" effect size (reduction of infarct volume by 48%, as assessed by meta-analysis), simulations showed that multi-laboratory studies based on as few as three laboratories can increase reproducibility from less than 50% to over 80%, without increasing false negative rate or a need for larger sample sizes[30].

Beyond HBA in ethical review of animal research, the 3Vs could also become instrumental for peer-review of grant applications and manuscripts submitted for publication. It is laudable that the NIH has recently updated its guidelines for how to evaluate research proposals by including assessment of scientific rigor (https://grants.nih.gov/reproducibility/index.htm), and that more and more journals are endorsing the UK NC3Rs ARRIVE guidelines (https://www.nc3rs.org.uk/arrive-guidelines). However, assessing scientific validity more systematically based on the 3Vs could help develop these initiatives further toward more powerful guidelines. As with the 3Rs, there is no need for a fixed checklist approach. Instead, funders deciding on the allocation of grant money, authorities licensing animal experiments, and editors evaluating manuscripts for publication could all define their own criteria for assessing each of the 3Vs in a way that appears most conducive to the kinds of decisions at their hands. Besides facilitating decision making, this would also enhance the scientific validity and reproducibility of findings from animal research. While this is clearly important for scientific reasons, it also matters on ethical grounds; it helps to avoid wasting animals for inconclusive research and imposing unnecessary harm on laboratory animals.

ACKNOWLEDGMENTS

I would like to thank Eimear Murphy, Katharina Friedli and Herwig Grimm for valuable comments on earlier drafts of this article. Research on which this article is based was funded by the European Research Council (ERC Advanced Grant REFINE No. 322576) and the Swiss Federal Food Safety and Veterinary Office (FSVO Grant No. 2.13.01).

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Bailoo, J.D., Reichlin, T.S. & Würbel, H. Refinement of experimental design and conduct in laboratory animal research. *ILAR J.* **55**, 383–391 (2014).
2. Brønstad, A. *et al.* Current concepts of harm–benefit analysis of animal experiments – report from the AALAS–FELASA working group on harm–benefit analysis – part 1. *Lab. Anim.* **50** 1S, 1–20 (2016).
3. Russell, W.M.S. & Burch, R.L. 1959. *The Principles of Humane Experimental Technique*. Methuen, London.
4. Vogt, L., Reichlin, T.S., Nathues, C. & Würbel, H. Authorization of animal experiments is based on confidence rather than evidence of scientific rigor. *PLoS Biol.* **14**, e2000598 (2016).
5. Ioannidis, J.P.A., Fanelli, D., Dunne, D.D. & Goodman, S.N. Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol.* **13**, e1002264 (2015).
6. Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
7. Macleod, M.R. *et al.* Risk of bias in reports of *in vivo* research: a focus for improvement. *PLoS Biol.* **13**, e1002273 (2015).
8. Reichlin, T.S., Vogt, L. & Würbel, H. The researchers' view of scientific rigor - Survey on the conduct and reporting of in vivo research. *PLoS ONE* **11**, e0165999 (2016).
9. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712 (2011).
10. Begley, C.G. & Ellis, L.M. Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
11. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
12. O'Collins, V.E. *et al.* 1,026 experimental treatments in acute stroke. *Ann. Neurol.* **59**, 467–477 (2006).
13. Nestler, E.J. & Hyman, S.E. Animal models of neuropsychiatric disorders. *Nat. Neurosci.* **13**, 1161–1169 (2010).
14. Button, K.S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
15. Richter, S.H., Garner, J.P. & Würbel, H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat. Methods* **6**, 257–261 (2009).
16. Head, M.L., Holman, L., Lanfear, R., Kahn, A.T. & Jennions, M.D. The extent and consequences of P-hacking in science. *PLoS Biol.* **13**, e1002106 (2015).
17. Tsilidis, K.K. *et al.* 2013. Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biol.* **11**: e1001609.
18. Sena, E.S., van der Worp, H.B., Bath, P.M.W., Howells, D.W. & Macleod, M.R. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol.* **8**, e1000344 (2010).
19. Bateson, P. When to experiment on animals. *New Sci.* **109**, 30–32 (1986).
20. van der Staay, F.J., Arndt, S.S. & Nordquist, R.E. Evaluation of animal models of neurobehavioral disorders. *Behav. Brain Funct.* **5**, 11 (2009).
21. Willner, P. Validation criteria for animal models of human mental disorders: learned helplessness as a paradigm case. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **10**, 677–690 (1986).
22. Van der Worp, H.B. *et al.* Can animal models of disease reliably inform human studies? *PLoS Med.* **7**, e1000245 (2010).
23. Paylor, R. Questioning standardization in science. *Nat. Methods* **6**, 253–254 (2009).
24. Richter, S.H., Garner, J.P., Auer, C., Kunert, J. & Würbel, H. Systematic variation improves reproducibility of animal experiments. *Nat. Methods* **7**, 167–168 (2010).
25. Richter, S.H. *et al.* Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS ONE* **6**, e16461 (2011).
26. Wodarski, R. *et al.* Cross-centre replication of suppressed burrowing behaviour as an ethologically relevant pain outcome measure in the rat: a prospective multicentre study. *Pain* **157**, 2350–2365 (2016).
27. Voelkl, B. & Würbel, H. Reproducibility crisis: Are we ignoring reaction norms? *Trends Pharmacol. Sci.* **37**, 509–510 (2016).
28. Parker, R.M.A. & Browne, W.J. The place of experimental design and statistics in the 3Rs. *ILAR J.* **55**, 477–485 (2014).
29. Würbel, H. Behaviour and the standardization fallacy. *Nat. Genet.* **26**, 263 (2000).
30. Würbel, H., Reichlin, T.S., Voelkl, B. & Vogt, L. 2016. More than refinement – improving the validity and reproducibility of animal research. in: Dwyer, C., Haskell, M., Sandilands, V. (eds.), *Proc. 50th Congr. Int. Soc. Appl. Ethol.*, Wageningen Academic Publishers, Wageningen, p. 324.