Animal Welfare 2019, 28: 107-115
ISSN 0962-7286
doi: 10.7120/09627286.28.1.107

How our approaches to assessing benefits and harms can be improved

ES Sena and GL Currie*

Centre for Clinical Brain Sciences, University of Edinburgh, Chancellor's Building, 49 Little France Crescent, Edinburgh EH16 4SB, UK
* Contact for correspondence and requests for reprints: emily.sena@ed.ac.uk

Abstract

Harm-benefit analysis (HBA) underpins the ethical framework of the regulation of animal experiments. This process involves a qualitative, and generally subjective, assessment of the potential benefits weighed against likely harms to be caused to animals. However, there is scope to prospectively quantify this process. A systematic and empirical assessment of historical data can give insights into why benefits are not realised and the magnitude of harm that animals experience. There is substantial scholarly evidence that risks to the 3Vs, the three core aspects of experimental validity in animal experiments (internal, external and construct validity) and low statistical power are limiting the reliability and reproducibility of research. Assessment of the 3Rs (reduction, refinement and replacement) is embedded in HBA and specifically seeks to minimise harm to the animals. However, no formal structure is in place to assess the likelihood of benefit, and we champion the 3Vs as a scale with which this may be achieved. Ethical approval procedures that consider the 3Vs and 3Rs using meta-research may be an approach to facilitate HBA. In ethical considerations related to animal research, there are value judgements that are integral to HBA, which cannot be measured directly. However, a quantitative and systematic approach is likely to be of added value. The perspective and examples described in this paper relate to laboratory animal research, but the approaches may lend themselves to different settings involving animals to ensure that decision-making and changes introduced, for example, to improve animal welfare, are evidence-based.

Keywords: *animal welfare, benefit, experimental validity, harm, laboratory research, meta-research*

Introduction

Research involving animals is performed for many reasons, including to understand biology, behaviour, conservation, diseases, inform the development of novel therapies, test the safety of agents for regulation and to understand how we may improve husbandry for animals kept in zoos, farms and laboratories. Societal interactions with animals extend beyond animal research, such as consuming them as food, keeping them as pets, in zoos or on farms, and exterminating those we consider pests. This paper will discuss the ethical framework we use to determine the appropriateness of animal use in research and how this may have broader implications for some societal interactions with animals.

Why do we undertake harm-benefit analysis (HBA)?

In many jurisdictions, harm-benefit analysis (HBA) is a legal instrument that underpins the ethical framework for the regulation of animal experimentation. This social and ethical evaluation applied to proposed research seeks to evaluate whether the harms that will be caused to protected animals, in terms of suffering, pain, distress, and lasting harm, can be justified by the expected benefit to humans, animals or the environment. The notion of benefits differs between different areas; in a zoo, benefits might relate to conservation of a species; in farming, there are benefits in increasing animal welfare and business viability, whilst reducing environmental damage. In the context of biomedical research, the expected benefits are to human beings. There is the expectation that the research will, for example, improve understanding of disease pathology, identify therapeutic targets and/or lead to the development or refinement of therapeutics.

In a 1986 article in the *New Scientist*, Patrick Bateson discussed the impending UK legislation to regulate the use of animals in research — the Animal (Scientific Procedures) Act 1986 — and the conflicts with those vehemently against the use of animals in research on moral grounds. He presented a decision cube (now known as Bateson's cube) as a framework to inform ethical considerations in decision-making processes of animal research (Bateson 1986). There are three 'sides' to the cube:

- The degree of animal suffering;
- The quality of the research;
- The benefits of the findings.

Ethical animal research should have clear benefits, be of high quality and ensure minimal suffering to the animal. Bateson's cube was not intended to provide a formal framework to assess trade-offs between the severity of harms and potential benefits because the axes of the cube are not in a common currency, therefore, we cannot balance these incommensurable properties (Bateson 2005). However, it was intended to inform sensible decision-making, such as in HBA.

The expectation is that HBA considers available current evidence of the harms to animals and potential benefit in a manner that is transparent, efficient and accountable; these evaluations should be performed by a sample of people that reflect societal concerns of the population. Whilst the evaluation of social and ethical issues is inherently subjective, it is evident that there is a robust and empirical scientific component that should underpin this undertaking. The 2017 Animal in Science's Committee review on HBAs in the UK addresses the current limitations in this process and the difficulties to ascertain the extent to which all current available evidence informs this process and how relevant new data are incorporated as, and when, they become available (Davies 2018). It may be that some committees undertaking HBAs do have a process that includes comprehensive and empirical up-to-date evidence of harms and likely benefit, but we are not of aware of any.

Our proposal seeks to apply an empirical assessment of experimental validity to assess the quality of research to weigh against the 3Rs (reduction, refinement and replacement) assessment which will inform an HBA. Whilst assessment of the 3Rs is already embedded in HBA and specifically seeks to minimise harm to the animals, no formal structure is in place to assess the quality of the research and thus the likelihood of benefit. This builds on Wuerbel's original description that the three core aspects of experimental validity in animal experiments (internal validity, external validity and construct validity; 3Vs) form the basis of such a scale as a counterpart to the 3Rs (Wuerbel 2017).

How may we take an empirical approach to provide added value to HBA?

Meta-research (research on research) provides a transparent and comprehensive approach to study how research is performed and interpreted. Its roots are in systematic review, but a broad range of methodologies may be employed. This approach applied to assess methods, reporting, evaluation, reproducibility and incentives, allows us to reach a rigorous understanding of what makes research reliable, and how it can be most effectively improved (Ioannidis *et al* 2015). Including such systematic and empirical approaches to HBA that are informed by the totality of relevant research, evidence may improve how we estimate the likelihood of benefit and severity of harm. The examples we present are primarily derived from biomedical research, but these concepts are just as relevant to veterinary research and may inform how we assess the welfare of animals in a non-research setting. Our aim here is to instigate discussion about how an empirical approach can aid HBA.

Benefit

For research to be of maximal value, and efficiently realise its intended benefit, it needs to be robust and of high quality. Currently, the likelihood of realising intended benefit is generally determined by assessment of the rigour of scientific questions being addressed by the intended research and the calibre of the scientist responsible for the research. In an academic setting, much of this evaluation is formally assessed prior to an HBA. The importance of a research question is assessed during the competitive grant awarding process. Because most research seeking ethical approval has already been deemed worthy for funding, this may influence ethical approval. This process is inextricably linked to the reputation of the lead scientist who is judged against criteria related to the quality of the journals in which they publish their research and their ability to secure competitive research funds. Unfortunately, there is no evidence that this assessment of academic success correlates with high experimental validity and reproducibility. We would argue that assessment of experimental validity offers a more appropriate correlate of the degree of purported benefit that is likely to be achieved. Laboratory animal research, particularly health-related research, has been blighted by positive findings observed in animal models that have not translated to similar effects in human studies (van der Worp *et al* 2010) and irreproducibility of findings between laboratories (Nosek & Errington 2017). Substantial meta-research has identified many limitations in the way in which we perform experiments using animal models of disease that likely contributes to this translational failure (Macleod *et al* 2015). We propose that an assessment of experimental validity should be an integral component of HBA.

Experimental validity refers to the extent to which variables influence the observed effects and the generalisability of these findings to other settings. There are various components to this validity that, if compromised, threaten the robustness of an experiment. Operationally, we define: 'internal validity' as the strength of the cause-effect relationship, eg whether the observed effects are due to the intervention rather than other unknown systematic biases; 'external validity' as the extent to which experimental inferences can be generalised, eg to other laboratories and/or, in the context of modelling human disease, to humans; 'construct validity' as the extent to which an outcome measure or experimental model measures what it purports to.

Systematic scrutiny of *in vivo* health research has allowed us to quantify how rarely, in some areas, the expected benefit is realised. In *in vivo* stroke research, of more than 1,000 interventions tested in animal studies, 596 were found to substantially improve outcome. Of these, 97 were tested in humans of which only one therapy was shown to be effective (O'Collins *et al* 2006). This level of attrition is high and the prior probability of developing an effective stroke therapy for humans is extremely low. Much meta-research has focused on the internal validity of experiments, and the presence and impact of potential threats to this validity. These include, but are not limited to, selection bias,

https://doi.org/10.7120/09627286.28.1.107 Published online by Cambridge University Press

performance bias, and detection bias which may be addressed by undertaking measures to reduce risks of bias, such as randomisation and blinding (Macleod *et al* 2015). Dissemination of these findings has been associated with improvements in the conduct and reporting of measures to reduce risks of bias in stroke experiments (McCann *et al* 2016; Minnerup *et al* 2016).

Assessment of internal validity should include whether appropriate controls have been used and measures to reduce the risk of bias have been undertaken. The internal validity of an experiment may be threatened by a range of biases. Selection bias occurs when there are systematic differences between study groups at the start of an experiment. Performance bias occurs when systematic differences occur in how the groups are handled during a study, and detection bias occurs when systematic differences occur between groups in how outcomes are ascertained, diagnosed, or verified. Measures to mitigate the risk of these biases include randomisation, allocation concealment and blinded (masked) assessment of outcome. Vogt and colleagues have demonstrated that, in Switzerland at least, few study protocols submitted for ethical evaluation report measures to reduce the risk of bias (2–19%) and corresponding publications report similarly low values (0–34%) (Vogt *et al* 2016). Similar analyses assessing the reporting of measures to reduce risks of bias in the published literature show equally high threats to internal validity (Macleod *et al* 2015). Across pre-clinical research domains, meta-research has demonstrated that studies that do report these measures to reduce risks of bias are associated with overestimated treatment effects (Crossley *et al* 2008; Macleod *et al* 2008; Vesterinen *et al* 2010; Rooke *et al* 2011; Hirst *et al* 2013).

External validity, our ability to replicate findings beyond the single laboratory to demonstrate the robustness of an effect, is essential (Wuerbel 2000). Limits to the external validity of pre-clinical research have been described as key contributors to the replication crisis and translational failure. To date, meta-research to improve the external validity of pre-clinical research has focused on the reporting of *in vivo* studies (eg characterisation of the presence and impact of reporting biases). Our predisposition in favour of 'positive' findings, particularly in academia, has led to an environment in which the direction and magnitude of findings are more likely rewarded than research studies performed to a high degree of experimental validity. Thus, neutral or conclusive findings in contrast to the alternative hypothesis carry little acclaim or reward. The major fallout of this preference has led to substantial publication bias in the life sciences. These biases lead to the totality of evidence overestimating treatment effects by about one-third (Sena *et al* 2010) and an over-representation of statistically significant results than would be expected (Tsilidis *et al* 2013). It is reasonable to expect that if animals are used in experiments then these data should be disseminated, irrespective of their findings, to contribute to our distillation of knowledge; experimental data that are never disseminated will not be able to do this. Commitment to disseminate research findings that use protected animals should be considered in HBA, akin to the requirement to disclose findings from interventional clinical trials (Moorthy *et al* 2015). The mechanism to implement such a requirement is not straightforward. At least in the UK, licence applications ask for a data dissemination strategy, but compliance is not formally assessed. Assessing compliance may facilitate the mitigation of publication bias. However, it is likely that different types of research require different strategies. For example, Kimmelman and colleagues (2014) propose distinguishing between exploratory and confirmatory research, which would provide scientists with the freedom to explore and innovate (exploratory research) in contrast to robust and reproducible outputs (confirmatory research). For confirmatory research, a more robust approach should require studies to be pre-registered in an animal study registry akin to those required for clinical trials. The complexities of such an endeavour for animal research have been investigated (Wieschowski *et al* 2016) and animal-specific platforms have recently been launched (preclinicaltrials.eu) or are in the pipeline (The German Federal Institute for Risk Assessment [BfR], personal communications 2018) providing an avenue for researchers to do this. However, this should not exempt researchers conducting exploratory research from a commitment to disseminate their findings, even if pre-registration is not appropriate.

Threats to external validity due to the way in which experiments are performed (ie experimental design) are also pertinent to the likelihood of realising benefit. For example, the role of experimental standardisation has generally been overlooked as a threat to external validity. Contrary to conventional wisdom that standardisation guarantees reproducibility (Beynen *et al* 2003) both theoretical (Wuerbel 2000; Voelkl & Wurbel 2016) and empirical (Crabbe *et al* 1999; Richter *et al* 2009, 2010; Kafkafi *et al* 2017) evidence indicates that rigorous standardisation may contribute to poor reproducibility. This is because the interaction between animal genotype and environmental conditions results in a specific phenotypic state that determines experimental response. The resulting range of response variation (the reaction norm) is seen as a nuisance that researchers often seek to eliminate through standardisation. Researchers can seek to formally sample from across the reaction norm by undertaking efforts to split experiments into multiple independent replicates (batches) (Paylor 2009), introducing systematic variation (systematic heterogeneity) (Richter *et al* 2010, 2011) of relevant variables (eg strains, housing conditions, tests, etc) or by implementing multi-centre study designs (Wodarski *et al* 2016; Voelkl *et al* 2018). Multi-centre studies allow us to sample across the reaction norm by utilising differences in environmental factors between laboratories, increasing generalisability and the likelihood of reproducibility. In some legal jurisdictions (eg Germany), animal experiments that address a research question that is deemed to have been answered already would not be granted ethical approval. It is unclear how the validity of such experiments is taken into account and how reproducible these findings may be when a research question is deemed to have been answered. A more prudent approach would be to assess

the experimental validity, including the generalisability, of studies addressing the same central hypothesis. At present, the lack of a structure to determine, or even consider, whether a central hypothesis has been robustly confirmed may cause problems for those wishing to undertake replication studies (or even multi-centre studies) to validate findings in an independent laboratory; an approach counter-intuitive to good science (Mogil & Macleod 2017).

In the context of pre-clinical studies of human diseases associated with co-morbidities (eg age, diabetes or hypertension in stroke) there is evidence that the majority of studies use young healthy male rats to model the disease and that when co-morbidities are introduced these experiments manifest smaller treatment effects (Sena *et al* 2007). Whilst it is not appropriate, nor desirable, that every *in vivo* study uses animals with co-morbidities or is multi-centre in its design, it is important to describe the next level of research that a study intends to inform and thus the intended external validity of the study. For example, in proof-of-principle studies where the intention is to inform the more complex experiments that model co-morbidities, the appropriate design, will differ from experiments that are intended to inform clinical trial design.

Modelling human diseases can be complex. Appropriate animal models and outcome measures are required to demonstrate comparable aetiology, and similar pathophysiology (face validity) and response to treatment (predictive validity) as observed in the human disease (Willner & Mitchell 2002). We describe construct validity as the experimental demonstration of these similarities. Assessment of construct validity should take into account the animal model, the test or outcome measure and the attribute it is intended to measure (Wuerbel 2017). Assessment of the tests and outcome measures should consider whether these are indeed measuring similar or independent constructs. Further, no one model can encapsulate the entirety of the heterogeneity observed in many diseases but understanding the construct validity of these models is important and providing evidence of face and predictive validity of these models, in the context of human diseases, is essential to underscore their appropriateness. For many diseases, a wide range of models have been established across many species. It has been reported that model selection is often based on the perceived aspect of the disease they are representing or of the mechanism of action of the intervention in question (Macrae 2011). However, as evidenced from this example of ischaemic stroke that describes many different models and variations of those models, it is unclear whether the panoply of available models and outcomes, and the consequent limits to their construct validity, are considered in this selection and the extent to which laboratory convention plays in this decision-making. Clear justification of model and outcome choice to ascertain construct validity could substantially improve our assessment of likely benefit.

Whilst limiting threats to experimental validity does not guarantee that benefit will be realised, it does provide a quantitative component related to the likelihood of success. This, alongside the other qualitative and somewhat subjective judgements we make of the likelihood of benefit and need, may improve how we approach the benefits arm of an HBA. Careful scrutiny of experimental validity using the 3V principles moves us from taking validity and reproducibility for granted (Wuerbel 2017) when assessing the likelihood of benefit.

Harm

Presently, in HBA, the likelihood of benefit is balanced against the estimated harm caused to animals. Our approach to assessing the harm to animals of proposed research is embedded in our application of the 3R (reduction, replacement and refinement) principles (Russell & Burch 1959). Many jurisdictions specifically require formal assessment of these principles and that those undertaking HBAs are satisfied that the research questions cannot be answered by using fewer animals (reduction), by using non-sentient animals or alternatives to *in vivo* models (replacement), and that animal welfare is maximised and less harmful procedures are not available (refinement). The implication being that this approach will ensure the suffering, pain, distress, and lasting harm to the animals is minimised as much as possible. Public opinion polls indicate that the majority of the public accept the use of animals in scientific research as long as there is no unnecessary suffering to the animals and there is no alternative (Leaman 2014).

The concept of reduction seeks "to reduce the numbers of animals used to obtain information of a given amount and precision." It is reasonable to argue that using too few animals to reliably answer a research question is unethical because these animals do not accurately contribute to our scientific knowledge. In fact, the National Centre for the Refinement, Replacement and Reduction of Animals in Research (NC3Rs) set up by the UK Government to oversee and promote laboratory animal welfare research, introduced a contemporary definition of reduction to reflect this "Appropriately designed and analysed animal experiments that are robust and reproducible, and truly add to the knowledge base." The appropriate number of animals required to answer a research question is an empirical question that can be determined by undertaking a sample size calculation based on estimates of variance, the desired magnitude of effect and the desired statistical power to detect this effect; yet, this estimation is rarely reported (Sena *et al* 2007; Macleod *et al* 2015). *Post hoc* calculations based on published data suggest that experiments should be substantially larger than reported in the literature. Indeed, the typical ischaemic stroke study is powered at only 30% — even if the effect being sought is present, there is a 70% chance that it will not be detected (CAMARADES, data on file). The reliability and accuracy with which conclusions are drawn from statistical tests are dependent on adequate sample size and the use of an appropriate statistical test for the type of data. Ioannidis has shown (Ioannidis 2005) that underpowered studies, resulting from inadequate sample sizes, are likely to have low positive predictive value, which is particularly relevant in laboratory research

https://doi.org/10.7120/09627286.28.1.107 Published online by Cambridge University Press

where statistical power is rarely considered (Sena *et al* 2007). Most animal studies are carried out under tightly defined conditions in healthy young animals (ie designed explicitly to show large effect sizes) and are underpowered to detect potentially important but smaller treatment effects. Studies seeking to detect smaller effects in more generalisable, less tightly controlled conditions in older animals with relevant co-morbidities are even more at risk of being underpowered (Festing & Altman 2002, 2005). Sample size calculations are most suitable for confirmatory research or exploratory research using inferential statistics.

When undertaking HBA, in addition to considering the design of the study, a meta-approach could be used to determine the importance of the research question. This may contribute to reducing the number of experiments conducted using animals.

The concept of refinement seeks to utilise methods that minimise pain, suffering, distress and lasting harm to animals and improve welfare. Given the panoply of approaches to animal experimentation and the ability to apply refinement approaches to all aspects of animal uses (eg from housing, husbandry, to the tests used) there are options to estimate the relative impact of pain and distress at these different stages of experimentation. Using a systematic approach, we can provide empirical evidence of the relative impact of experimental design choices on measures of animal welfare. For example, environmental enrichment has been widely advocated, in laboratory and zoo settings, to improve the welfare and the quality of research (Baumans 2005). An added potential benefit of environmental enrichment is the increased environmental heterogeneity within a study, thus increasing the sampling of the reaction norm, for improved external validity. However, in a systematic review describing the modelling of chemotherapy-induced peripheral neuropathy, only five of 341 publications (1%) reported home-cage enrichment (Currie *et al* 2018). An additional consideration of animal welfare in the modelling of pain is post-operative analgesia. There is no consensus on the effect of post-operative analgesia regimens on experimental outcomes and in experiments using animal models of neuropathic pain often only minimal analgesics are given. There is an ongoing meta-research project to compare the effect of different post-operative analgesia regimens on modelling (Currie *et al* 2018; http://www.dcn.ed.ac.uk/camarades/research.html#protocols) to provide empirical evidence to support these decisions. This project may support the hypothesis that robust analgesics can be administered to animals and only ceased for a window when measurements are taken. This could potentially reduce the duration of suffering that animals experience during pain-modelling experiments. In the modelling of spinal cord injury, by comparing findings from studies where different outcomes have been measured in the same cohort of animals, there is evidence that subjecting animals to additional distress, by subjecting them to multiple tests, provided no added value in terms of demonstrating effectiveness (Antonic *et al* 2013). Providing empirical evidence to determine whether less-noxious tests are as predictive as more severe alternatives and whether multiple tests are necessary would be a useful addition to the implementation of refinement strategies.

Replacement seeks to determine whether it is possible to avoid or replace the use of animals in research. This may refer to the use of non-*in vivo* methods, such as *in vitro* or computational models. Further, relative replacement sets out to answer the same research question using non-sentient animals or in early-life stages before legal protection exists. When researchers justify the use of animals in research, it is not immediately apparent how this is formally assessed beyond statements from researchers that alternatives are not appropriate to address their research question. Central resources to identify alternatives are limited, we are only aware of a German-speaking one (The German Centre for Documentation and Evaluation of Alternatives to Animal Experiments [ZEBET]), and few researchers formally search for alternatives in the design of their experiments (van Luijk *et al* 2011). The NC3Rs 2017–2019 strategy discusses the 3Rs 'valley of death', the gap between the development of new 3Rs technologies and approaches and their adoption into routine use. They have recently launched a platform in collaboration with F1000 to address this problem (Percie du Sert & Robinson 2018). However, this platform is only for NC3Rs-funded research meaning that 3Rs-relevant research not funded by the NC3Rs will not contribute to conquering this valley of death. Further, relevant research that does not explicitly highlight the 3Rs opportunities compounds this problem. A formal requirement in license applications to detail efforts made to identify 3Rs opportunities could make the development of a more comprehensive tool more likely and highly impactful. In addition, a strategy to better reward the legacy of 3Rs developments may facilitate more effective implementation of 3Rs approaches.

Practical applications

Providing an empirical component to HBA that considers both the 3Rs and 3Vs underscores the link between good animal welfare and high-quality science. There is an inherent reluctance to question and challenge established practices and cultures, and even where evidence exists to improve animal welfare or increase experimental validity, routine or uniform implementation does not yet exist. A systematic overview of research areas to understand the gaps that intended research seeks to address would be useful, but resource and skills to undertake such research are required. It is not reasonable to envisage that *in vivo* researchers alone should shoulder such responsibility; meta-research is a fast-developing discipline that is further complicated by the high rate with which we accrue data and studies are being published.

A formal framework to assess the 3Rs and 3Vs requires a nuanced approach that takes into consideration the stage in the research pipeline for which ethical approval is being sought and how it fits into the broader research landscape of a research area. For example, in exploratory research seeking to test the effectiveness of a novel treatment in a model of human disease, a framework to assess the external validity is complex. It may be that this treatment is being developed in a single laboratory, and whilst the external

validity of such a study is likely to be restricted, due to limited sampling across the reaction norm, it is important that the limits to the generalisability of the proposed study are considered in the context of the proposed research question. This consideration should be described at all stages of the research lifecycle, from proposals submitted to funders, ethical approval applications and any resulting publications or other modes of dissemination of research findings. However, if ethical approval is being sought for a confirmatory study; one that seeks to determine the effectiveness of this intervention to inform clinical trial design, the research design should be externally valid. We envisage that such a study samples appropriately from the reaction norm, or that the desire to replicate a study, for which an effect has been observed in a single centre exploratory study, to demonstrate external validity is not penalised. To perform an informed assessment of the likelihood of benefit, particularly in terms of the external validity, a systematic review that details the range of conditions that the intervention has demonstrated effectiveness is required. An assessment of the use of measures to reduce risks of bias, to ensure internal validity, and the limits to the construct validity in terms of models and outcome measures that have been sampled are also required. This is a substantial undertaking and would require investment of resources to perform such activities that funding agencies should consider as due diligence prior to further investment of larger more resource-intense research studies. We do not advocate that animal research should only be performed in light of a systematic review of relevant evidence. However, when considering the undertaking of large, potentially multi-centre, confirmatory studies, it would be prudent to ensure the premise of such a study is robust. We call for more meta-research, on which researchers and ethics committees can draw upon in their decision-making.

There is scope to employ a scoring system of pre-defined measures to assess internal validity, including plans for randomisation and blinding, across the life sciences. There may be cases where blinding or randomisation are not feasible (eg fur colour differs between cohorts) and this should be considered. Further, findings should be reported irrespective of the direction of effect to ensure that the animals used contribute to our accumulation of knowledge. To protect against reporting biases, particularly in confirmatory research, it would be appropriate for the investigators to assure *a priori* registration of their study protocol which details the experimental design, planned outcomes and analysis plan. Advocated by the Centre for Science (COS), registered reports are a type of publication that seeks to address many of these issues. The concept revolves around the quality of a research question and methodology by peer review of an introduction and methods prior to any data collection. High-quality protocols, which detail experimental design, planned outcomes and analysis plan are registered and the completed study is provisionally accepted for publication if the authors follow the registered methodology (Chambers *et al* 2014).

Unlike the 3Vs, formal assessment of the 3Rs is already embedded in our assessment of HBA but much of this appears to be a qualitative undertaking. There is scope to introduce a more systematic approach to understanding potential 3Rs options. For example, this may include empirical and comprehensive evidence of the impact of housing and husbandry choices on measures of welfare. Further, as 3Rs alternatives become available, it is important that approaches for their dissemination are fit-for-purpose and that the validity of this evidence is robust. The NC3Rs/F1000 portal seeks to address this for NC3Rs-funded researchers but dissemination strategies beyond this cohort of researchers are required. For reduction, which may be approached in an empirical manner, particularly in confirmatory research, sample size calculations ought to be performed. As efforts continue to develop novel methods to measure welfare or to develop alternatives to animal research, it is important that systematic investigation continues to cycle through testing and evaluation of recommendations.

HBAs are a prospective analysis but there is opportunity to perform retrospective HBA analyses that have the potential to improve substantially the HBA process. Since 2014, European legislation requires the recording of actual harms suffered by individual animals with the view to refine severity categories of procedures for prospective analyses. Expanding such retrospective analyses to include assessment of benefits, through meta-research, has the potential to inform retrospective HBA analysis. A recent study (Pound & Nicol 2018) sought to operationalise retrospective HBA using Bateson's cube as a framework. Their approach was somewhat different to our proposal here but is the first time, as far as we are aware, systematic retrospective HBA has been attempted. Even in research areas where there was high concordance between animal and clinical findings, and benefit has been realised in terms of clinical use, none of the 212 studies were considered ethical using this framework. To determine this, 212 studies were categorised by severity of harms to animals and the benefit was assessed in terms of clinical relevance. Pound and Nicol concluded that many animals suffered severe harms that were not associated with benefits for humans and only a small proportion of studies minimised harms to animals whilst being associated with human benefit.

In considering benefit it is important to acknowledge the mechanism with which the findings from animal research is assessed and how the potential benefits are implemented at the next stage of research. Systematic approaches that focus on the validity of research findings have been proposed to determine the certainty of evidence of pre-clinical animal research and whether sufficient high-quality evidence exists prior to embarking upon human clinical trial (Hooijmans *et al* 2018). This Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach uses a framework that includes aspects of the 3Vs in this assessment. It is also important to consider that downstream research in humans that is informed by animal research (phase I/II clinical trials) undergoes a risk-harm analysis during the ethical approval process that is often

https://doi.org/10.7120/09627286.28.1.107 Published online by Cambridge University Press

informed by pre-clinical research. In a study of investigator brochures that are submitted for ethics review, the quality of the animal research cited was limited (Wieschowski *et al* 2018). Less than 5% of the animal studies cited reported a sample size calculation, randomisation and blinding; 89% were not published and the majority (82%) only described positive finding. A similar approach to understanding the likelihood of benefit based on the 3Vs would be a useful approach in this context.

Animal welfare implications

The HBA process could be improved by the addition of empirical assessment of the 3Rs and the 3Vs. By quantifying these factors, we can evaluate whether harms are justified and that harms are minimised and balanced against the likelihood of benefit. An important animal welfare implication is that such a meta-approach allows for an evidence-based process to improve animal welfare and underscores a framework to prevent animals being subjected to harms where the benefits are unlikely to be reached. Formal application of HBA is only required for animals protected by legislation but the ethical consideration of harms and benefits are applied more broadly in our societal interaction with animals. This is not just an issue for laboratory science, there are many settings where animal welfare is important and could benefit from a systematic approach to assessing welfare and the totality of evidence that could be applied to decisions; for example, environmental enrichment approaches for farm or zoo animals. There is scope to undertake meta-approaches across different settings to determine consistency across settings and species.

Conclusion

In considering Bateson's three axes, our assessment of 3Rs and 3Vs focuses on (i) the degree of animal suffering and (ii) the quality of the research. The third axis, (iii) benefit of the findings, relies on the unmet need addressed by the research question. In this paper, this has focused on unmet medical need (eg therapies for stroke or chronic pain). There is clear consensus that animal research should have evident benefits, be of high quality and cause minimal suffering to the animal. However, the mode with which to operationalise this is less clear. Undertaking meta-research to support study protocols considered in HBA will provide a robust framework to assess both the 3Rs and the 3Vs. This approach would ensure that the entirety of relevant literature is considered when estimating the magnitude of harms that may be inflicted and that changes introduced are evidence-based. Further, by considering the intended beneficiaries of an experiment and quantifying the experimental validity, it may be that these benefits will be substantially more realistic.

Focusing on the 3Vs provides the capacity to ensure that a study is robust and contributes to our distillation of knowledge, which is likely to be in small increments rather than a large unsubstantiated claim of a benefit to humanity that is seldom realised from a single study. This approach hopefully will increase levels of trust in the research conducted and ensure the validity of the knowledge gained from these experiments.

Acknowledgements

We are grateful to Dan Weary, Michael Appleby and Pete Sandøe, and the Wissenschaftskolleg zu Berlin, Institute for Advanced Study for hosting and organising the stimulating 'Animal Welfare Reconsidered' workshop, and to all participants of the workshop and Kaitlyn Hair for their insightful comments and discussions of earlier versions of this manuscript. This study was supported by the UK National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) infrastructure award: ivSyRMAF — the CAMARADES — NC3Rs *in vivo* systematic review and meta-analysis facility (NC/L000970/1) and the Stroke Association (SA L-SNC 18\1003). The study sponsors had no role our interpretations, writing of the report or decision to submit the manuscript for publication.

References

Antonic A, Sena ES, Lees JS, Wills TE, Skeers P, Batchelor PE, Macleod MR and Howells DW 2013 Stem cell transplantation in traumatic spinal cord injury: a systematic review and meta-analysis of animal studies. *PLoS Biology 11*: e1001738. https://doi.org/10.1371/journal.pbio.1001738

Bateson P 1986 When to experiment on animals. *New Scientist 109*: 30-32

Bateson P 2005 Ethics and behavioral biology. *Advances in the Study of Behaviour* pp 211-233. http://www.psych.utoronto.ca/users/psy3001/files/Bateson%202005.pdf

Baumans V 2005 Environmental enrichment for laboratory rodents and rabbits: requirements of rodents, rabbits, and research. *ILAR Journal 46*: 162-170. https://doi.org/10.1093/ilar.46.2.162

Beynen A, Gärtner K and van Zutphen L 2003 Standardisation of animal experimentation. In: Zutphen L, Baumans V and Beynen A (eds) *Principles of Laboratory Animal Science* pp 103-110. Elsevier: Amsterdam, The Netherlands

Chambers CD, Feredoes E, Muthukumaraswamy S and Etchells P 2014 Instead of 'playing the game' it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience 1*: 4-17. https://doi.org/10.3934/Neuroscience.2014.1.4

Crabbe JC, Wahlsten D and Dudek BC 1999 Genetics of mouse behavior: interactions with laboratory environment. *Science 284*: 1670-1672. https://doi.org/10.1126/science.284.5420.1670

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PMW, Macleod M and Dirnagl U 2008 Empirical evidence of bias in the design of experimental stroke studies: A metaepidemiologic approach. *Stroke 39*: 929-934. https://doi.org/10.1161/STROKEAHA.107.498725

Currie GL, Angel-Scott H, Colvin L, Cramond F, Hair K, Khandoker L, Liao J, Macleod MR, McCann SK, Morland R, Sherratt N, Stewart R, Tanriver-Ayder E, Thomas J, Wang Q, Wodarski R, Xiong R, Rice ASC and Sena ES 2018 Animal models of chemotherapy-induced peripheral neuropathy: a machine-assisted systematic review and meta-analysis A comprehensive summary of the field to inform robust experimental design. *bioRxiv 293480*. https://doi.org/10.1101/293480

Davies GF 2018 Harm-benefit analysis: opportunities for enhancing ethical review in animal research. *Laboratory Animals (NY) 47*: 57-58. https://doi.org/10.1038/s41684-018-0002-2

Festing MFW and Altman DG 2002 Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal/National Research Council, Institute of Laboratory Animal Resources 43*: 244-258. https://doi.org/10.1093/ilar.43.4.244

Festing MFW and Altman DG 2005 Guidelines for the design and statistical analysis of experiments using laboratory aminals. *ILAR Journal/National Research Council, Institute of Laboratory Animal Resources 43*: 244-258. https://doi.org/10.1093/ilar.43.4.244

Hirst T, Vesterinen H, Sena E, Egan K, Macleod M and Whittle I 2013 Systematic review and meta-analysis of temozolomide in animal models of glioma: was clinical efficacy predicted? *British Journal of Cancer 108*: 64-71. https://doi.org/10.1038/bjc.2012.504

Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M, Rovers MM, Leeflang MM, IntHout J, Wever KE, Hooft L, de Beer H, Kuijpers T, Macleod MR, Sena ES, ter Riet G, Morgan RL, Thayer KA, Rooney AA, Guyatt GH, Schünemann HJ, Langendam MW and on behalf of the GWG 2018 Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS One 13*: e0187271

Ioannidis JPA 2005 Why most published research findings are false. *PLoS Medicine 2*: 696-701. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis JPA, Fanelli D, Dunne DD and Goodman SN 2015 Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology 13*: e1002264. https://doi.org/10.1371/journal.pbio.1002264

Kafkafi N, Golani I, Jaljuli I, Morgan H, Sarig T, Wurbel H, Yaacoby S and Benjamini Y 2017 Addressing reproducibility in single-laboratory phenotyping experiments. *Nature Methods 14*: 462-464. https://doi.org/10.1038/nmeth.4259

Kimmelman J, Mogil JS and Dirnagl U 2014 Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology 12(5)*: e1001863. https://doi.org/10.1371/journal.pbio.1001863

Leaman J 2014 Attitudes to animal research in 2014. *A report by Ipsos MORI for the Department for Business, Innovation & Skills*. https://www.ipsos.com/sites/default/files/migrations/en-uk/files/Assets/Docs/Polls/sri_BISanimalresearch_TRENDreport.pdf

Macleod MR, Lawson MA, Kyriakopoulou A, Serghiou S, de WA, Sherratt N, Hirst T, Hemblade R, Bahor Z, Nunes-Fonseca C, Potluru A, Thomson A, Baginskaite J, Egan K, Vesterinen H, Currie GL, Churilov L, Howells DW and Sena ES 2015 Risk of bias in reports of *in vivo* research: A focus for improvement. *PLoS Biology 13*: e1002273

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U and Donnan GA 2008 Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke 39*: 2824-2829. https://doi.org/10.1161/STROKEAHA.108.515957

Macrae IM 2011 Preclinical stroke research: advantages and disadvantages of the most common rodent models of focal ischaemia. *British Journal of Pharmacology 164*: 1062-1078. https://doi.org/10.1111/j.1476-5381.2011.01398.x

McCann SK, Cramond F, Macleod MR and Sena ES 2016 Systematic review and meta-analysis of the efficacy of interleukin-1 receptor antagonist in animal models of stroke: an update. *Translational Stroke Research 7*: 395-406. https://doi.org/10.1007/s12975-016-0489-z

Minnerup J, Zentsch V, Schmidt A, Fisher M and Schabitz WR 2016 Methodological quality of experimental stroke studies published in the stroke journal: time trends and effect of the basic science checklist. *Stroke 47*: 267-272. https://doi.org/10.1161/STROKEAHA.115.011695

Mogil JS and Macleod MR 2017 No publication without confirmation. *Nature 542*: 409-411. https://doi.org/10.1038/542409a

Moorthy VS, Karam G, Vannice KS and Kieny MP 2015 Rationale for WHO's new position calling for prompt reporting and public disclosure of interventional clinical trial results. *PLoS Medicine 12*: e1001819. https://doi.org/10.1371/journal.pmed.1001819

Nosek BA and Errington TM 2017 Making sense of replications. *eLife 6*: e23383. https://doi.org/10.7554/eLife.23383

O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH and Howells DW 2006 1,026 experimental treatments in acute stroke. *Annals of Neurology 59*: 467-477. https://doi.org/10.1002/ana.20741

Paylor R 2009 Questioning standardization in science. *Nature Methods 6*: 253-254. https://doi.org/10.1038/nmeth0409-253

Percie du Sert N and Robinson V 2018 The NC3Rs gateway: Accelerating scientific discoveries with new 3Rs models and technologies. *F1000Research 7*: 591. Phttps://doi.org/10.12688/f1000research.14964.1

Pound P and Nicol CJ 2018 Retrospective harm benefit analysis of pre-clinical animal research for six treatment interventions. *PLoS One 13*: e0193758. https://doi.org/10.1371/journal.pone.0193758

Richter SH, Garner JP, Auer C, Kunert J and Wuerbel H 2010 Systematic variation improves reproducibility of animal experiments. *Nature Methods 7*: 167-168

Richter SH, Garner JP and Wuerbel H 2009 Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nature 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 SN, van der Harst J, Spruijt B, Voikar V, Wolfer DP and Wurbel H 2011 Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS One 6*: e16461

Rooke ED, Vesterinen HM, Sena ES, Egan KJ and Macleod MR 2011 Dopamine agonists in animal models of Parkinson's disease: A systematic review and meta-analysis. *Parkinsonism & Related Disorders 17*: 313-320. https://doi.org/10.1016/j.parkreldis.2011.02.010

Russell WMS and Burch RL 1959 *The Principles of Humane Experimental Technique*. Methuen: London, UK

Sena E, van der Worp HB, Howells D and Macleod M 2007 How can we improve the pre-clinical development of drugs for stroke? *Trends in Neurosciences 30*: 433-439. https://doi.org/10.1016/j.tins.2007.06.009

Sena ES, van der Worp HB, Bath PMW, Howells DW and Macleod MR 2010 Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biology 8*: e1000344. https://doi.org/10.1371/journal.pbio.1000344

https://doi.org/10.7120/09627286.28.1.107 Published online by Cambridge University Press

Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, Salman RA-S, Macleod MR and Ioannidis JP 2013 Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biology 11(7)*: e1001609. https://doi.org/10.1371/journal.pbio.1001609

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V and Macleod MR 2010 Can animal models of disease reliably inform human studies? *PLoS Medicine 7*: e1000245. https://doi.org/10.1371/journal.pmed.1000245

van Luijk J, Cuijpers Y, van der Vaart L, Leenaars M and Ritskes-Hoitinga M 2011 Assessing the search for information on Three Rs methods, and their subsequent implementation: a national survey among scientists in the Netherlands. *Alternatives to Lab Animals 39*: 429-447

Vesterinen HM, Sena ES, French-Constant C, Williams A, Chandran S and Macleod MR 2010 Improving the translational hit of experimental treatments in multiple sclerosis. *Multiple Sclerosis 16*: 1044-1055. https://doi.org/10.1177/1352458510379612

Voelkl B, Vogt L, Sena ES and Wuerbel H 2018 Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biology 16*: e2003693. https://doi.org/10.1371/journal.pbio.2003693

Voelkl B and Wurbel H 2016 Reproducibility crisis: Are we Ignoring reaction norms? *Trends Pharmacological Sciences 37*: 509-510. https://doi.org/10.1016/j.tips.2016.05.003

Vogt L, Reichlin TS, Nathues C and Wurbel H 2016 Authorization of animal experiments is based on confidence rather than evidence of scientific rigor. *PLoS Biology 14*: e2000598. https://doi.org/10.1371/journal.pbio.2000598

Wieschowski S, Chin WWL, Federico C, Sievers S, Kimmelman J and Strech D 2018 Preclinical efficacy studies in investigator brochures: Do they enable risk-benefit assessment? *PLoS Biology 16*: e2004879. https://doi.org/10.1371/journal.pbio.2004879

Wieschowski S, Silva DS and Strech D 2016 Animal study registries: Results from a stakeholder analysis on potential strengths, weaknesses, facilitators, and barriers. *PLoS Biology 14*: e2000391. https://doi.org/10.1371/journal.pbio.2000391

Willner P and Mitchell PJ 2002 The validity of animal models of predisposition to depression. *Behavioural Pharmacology 13*: 169-188. https://doi.org/10.1097/00008877-200205000-00001

Wodarski R, Delaney A, Ultenius C, Morland R, Andrews N, Baastrup C, Bryden LA, Caspani O, Christoph T, Gardiner NJ, Huang W, Kennedy JD, Koyama S, Li D, Ligocki M, Lindsten A, Machin I, Pekcec A, Robens A, Rotariu SM, Vob S, Segerdahl M, Stenfors C, Svensson CI, Treede RD, Uto K, Yamamoto K, Rutten K and Rice AS 2016 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. https://doi.org/10.1097/j.pain.0000000000000657

Wuerbel H 2000 Behaviour and the standardization fallacy. *Nature Genetics 26*: 263. https://doi.org/10.1038/81541

Wuerbel H 2017 More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. *Laboratory Animals (NY) 46*: 164-166. https://doi.org/10.1038/laban.1220