Bench Marks!

On the Usefulness of Animals as a Model System (Part I): Overview of Criteria and Focus on Robustness

Giorgia Pallocca[1], *Costanza Rovida*[1] *and Marcel Leist*[1,2]

[1]CAAT-Europe, University of Konstanz, Konstanz, Germany; [2]In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Konstanz, Germany

Abstract

Banning or reduction of the use of animals for laboratory experiments is a frequently-discussed societal and scientific issue. Moreover, the usefulness of animals needs to be considered in any decision process on the permission of specific animal studies. This complex issue is often simplified and generalized in the media around the question, "Are animals useful as a model?" To render an often emotional discussion about animal experimentation more rational, it is important to define "usefulness" in a structured and transparent way. To achieve such a goal, many sub-questions need to be asked, and the following aspects require clarification: (i) consistency of animal-derived data (robustness of the model system); (ii) scientific domain investigated (e.g., toxicology vs disease modelling vs therapy); (iii) measurement unit for "benefit" (integrating positive and negative aspects); (iv) benchmarking to alternatives; (v) definition of success criteria (how good is good enough); (vi) the procedure to assess benefit and necessity. This series of articles discusses the overall benchmarking process by specifying the six issues. The goal is to provide guidance on what needs to be clarified in scientific and political discussions. This framework should help in the future to structure available information, to identify and fill information gaps, and to arrive at rational decisions in various sub-fields of animal use. In part I of the series, we focus on the robustness of animal models. This describes the capacity of models to produce the same output/response when faced with the "same" input. Follow-up articles will cover the remaining usefulness aspects.

1 Setting the scene

All important technologies, procedures, materials, and processes need re-evaluation from time to time. Even those that changed the world, dominated the market, or greatly advanced civilization and/or knowledge may at some point need an update. Alternatively, they may become obsolete, be discontinued, or be relegated to niche applications; sometimes they are merged with other technologies to result in something new. Steam engines, electric lightbulbs, typewriters, leather skiing boots, the space shuttle program, CD drives on computers, cathode-ray tube monitors, and basic mobile phones without touch displays had such fates. In pharmacology and agriculture, some inhibitors of cyclooxygenase-2 or particular classes of pesticides are prominent examples of usefulness re-evaluation.

Some fields have developed their own culture of such evaluations, and important principles have been established for the judgement procedure. In medicine, it has become clear that "usefulness" is a multi-dimensional issue (Fig. 1): A first approach to judge drugs would address their "efficacy". Do they activate/block the desired target; how potently and how effectively? And does this lead to a desired effect (change of a disease symptom

Received March 29, 2022;
© The Authors, 2022.

ALTEX 39(2), 347-353. doi:10.14573/altex.2203291

Correspondence: Marcel Leist, PhD
In vitro Toxicology and Biomedicine
Dept inaugurated by the Doerenkamp-Zbinden foundation
at the University of Konstanz
Universitaetsstr. 10, 78464 Konstanz, Germany
(marcel.leist@uni-konstanz.de)

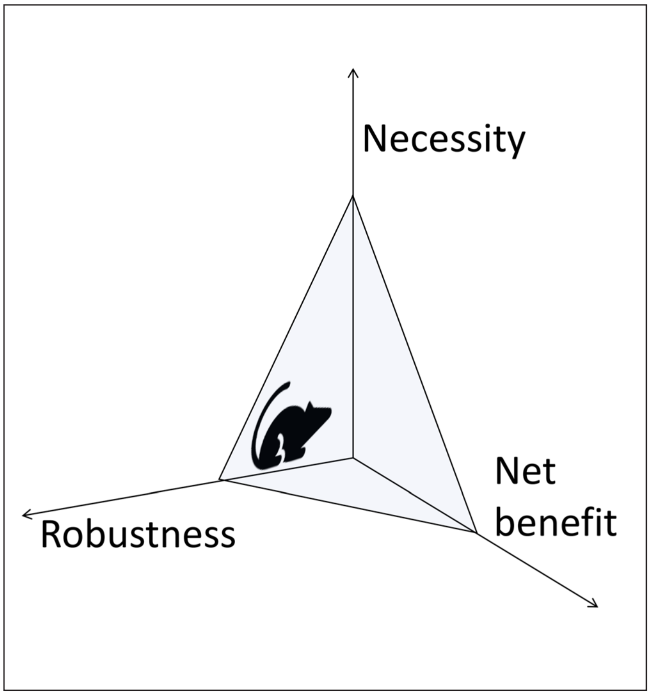

Fig. 1: Three key aspects of a test methods' practical usefulness
The usefulness of a test method depends on several aspects. There are at least three orthogonal (independent) aspects that affect the usefulness of any given application. Thus, several dimensions need to be considered when evaluating usefulness (e.g., of animal-based test methods). A method can be more or less robust (reliable in its output). At any given robustness level, the model has advantages and disadvantages. This ratio results in an overall "net benefit" (which can be negative or positive). Methods of a given net benefit level may be useful in a real-world setting in competition with other methods to solve the same question. The "necessity level" is the relative advantage (or disadvantage) compared to other methods.

or a pathophysiologic parameter)? On the next level, one would ask about "benefit". Two aspects distinguish benefit from efficacy: (i) modification of a target (or of a measurable efficacy parameter) may not benefit a diseased patient (in terms of quality of life, disease progression or severity of symptoms): (ii) concomitant adverse effects may reduce the benefit. Thus, the sum of all advantages and disadvantages for the patient must be considered. A third dimension goes beyond the evaluation of the drug (effects) as such to consider a wider context. This is often termed "necessity" or "comparative benefit". For example, which other drugs are available to treat the same disease/symptom? If there are alternatives that are more beneficial or cheaper or more readily available, the necessity for the new drug will be considered low; if no alternative is available, even a small benefit can be sufficient to deem a new drug beneficial. An important implication is that the necessity, i.e., the overall value of a drug, can change with time. Even a good (beneficial) drug loses necessity if a better one is found. An example is the acid blocker cimetidine, which was the best-selling drug worldwide for a while, but now is only used rarely.

The above example on drugs suggests that asking simply for the "usefulness of animals as a model" may not consider the multi-dimensionality of the problem. When the value of a "model" is at question, even more dimensions may need to be considered than for a drug (e.g., the modelling purpose).

There is a saying that there are no stupid questions. This may be true, but there are questions that are not helpful or that cannot be answered in a meaningful way. Sometimes, it is essential for a good discussion to ask the right questions. Examples for questions of little use are: Are potatoes healthy? Are bicycles fast? Is London well-situated? Do vitamins help? Are diesel engines useful? We feel that the question "Are animals useful as a model?" falls into this category, but the desire and motivation to learn more about the value of animal models are well justified. How do we resolve this apparent conflict? Our suggestion is to better specify the question and to give space to all the different layers of information and dimensions of reasoning that are essential to form an opinion on the overall issue.

2 What is needed to judge usefulness

"Usefulness" is a complex concept. In addition, the term "model" has many aspects. Thus, the question about the "usefulness of animals as a model" cannot be answered with yes or no, and it cannot be defined in two sentences. In this paragraph, we want to outline the landscape of approaches used to come to an evaluation (Fig. 2).

As a starting point, one may consider the features of toxicological models examined in a classical validation process: relevance, predictivity, and robustness (Krebs et al., 2020; Hartung et al., 2013; Judson et al., 2013; Hartung, 2007, 2010).

Robustness is the easiest to handle, especially if expressed in its reverse form: something that is not robust is clearly not useful. Robustness is sometimes also called reliability or reproducibility. It means that the model will produce the same output/response when faced with the "same" input (Leist et al., 2008, 2014). We discuss robustness in more detail in the next section.

Predictivity is generally meant to describe how well the results of the model correlate with reality or with a gold standard. Historical data are used to measure predictivity, and it is hypothesized that the same correlation will apply in the future. Obviously, the quality of prediction will depend on the calibration (often called training) of the model. It will also be affected by how well the future situation correlates with past situations of training. Thus, the development of scientific fields, the application domains of a model, and the objectives/goals of its use play a major role.

Relevance has been the most difficult to tackle in validation processes. It has several aspects, and one of the main approaches is to examine whether the internal working and its mechanisms are similar between model and reality. Sometimes relevance has been called "external validity" to distinguish it from internal va-

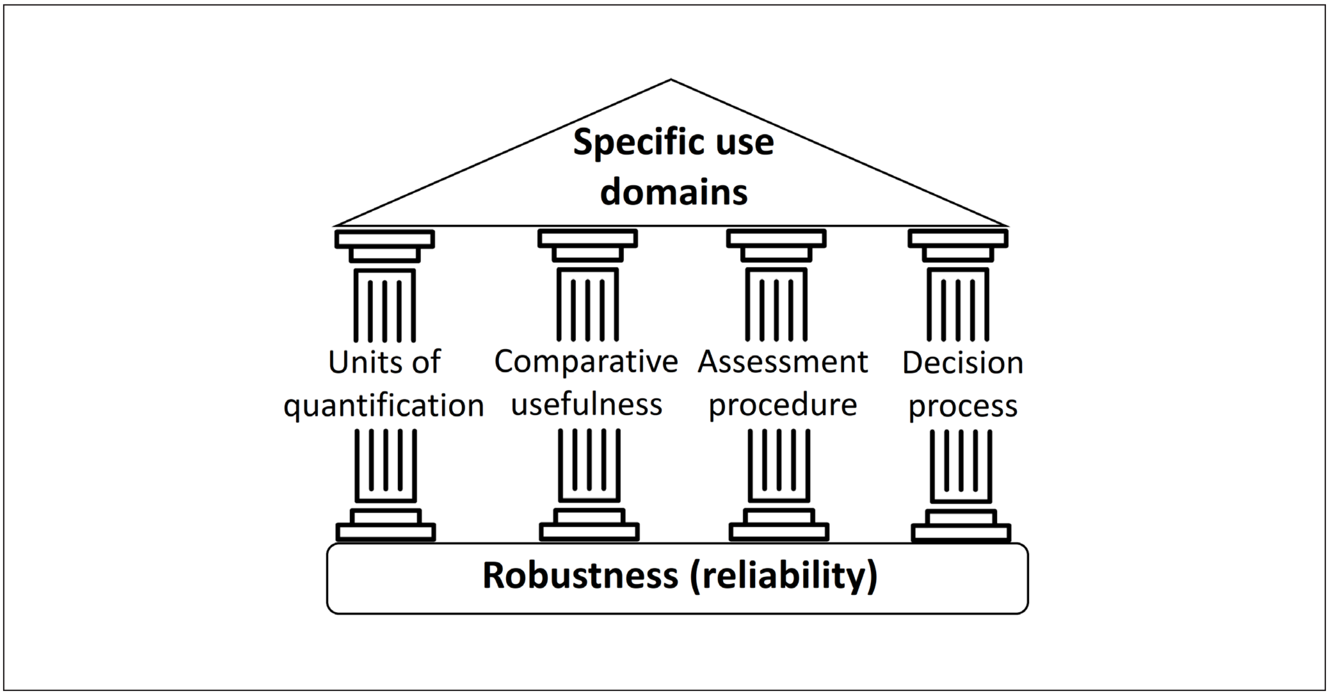

Fig. 2: "Building blocks" of test method usefulness assessment
The most important concept promoted here is that asking whether animal models are useful or not in general is a problematic approach. We rather suggest that each specific intended use should be considered and evaluated independently. The usefulness of each application domain forms a self-sufficient building. Here we indicate the objective as a roof structure supported by four important columns. These again need to stand on a foundation of robustness. The four supporting columns (i.e., tiers of usefulness evaluation) are the net benefit (sum of advantages and disadvantages), the necessity (the competitiveness vs alternative approaches), the method used for quantification, and the choice of the decision process. The latter includes, e.g., the weight given to timing aspects, i.e., which time period is evaluated for usefulness. The consequence of this approach could be that some domains of animal use turn out to be solid, granite-built fortresses, while others are shaky sheds that crumble at the first touch.

lidity, i.e., robustness (Ferreira et al., 2020). Quantification of this aspect is difficult. For instance, both rodents and humans show chemical-induced carcinogenesis, and many mechanisms leading to neoplasia are similar. This example would suggest a high score for relevance of rodent models for human carcinogenesis. On the other hand, the ratios of sarcomas and carcinomas largely differ between the species, as do the target organs, and the role of immune control. This aspect would score low on relevance. Similarly, occlusion of a cerebral artery in a rodent is similar to occlusion by a thrombus in humans (apparently high relevance), but the outcomes of pharmacological experiments in this situation are largely different (apparently low relevance). Several overviews are available that show that the relevance of animals for humans in biomedical and drug research is based on very thin data (Ferreira et al., 2020; Pistollato et al., 2020). For lack of better tools, "relevance" is often used as a proxy to quantify "predictivity". This assumption is, strictly speaking, not correct, as there may be models with apparently low relevance that show a high predictivity and *vice versa*.

A common, widespread fallacy in the validation process is the following: If a model has been found (empirically) to be useful for one study, it is assumed that it is useful in general and for any question. The statistician George Box brought forward the famous quote, "*All models are wrong, but some are useful*". This implies that usefulness may be related to concrete purposes and applications, and that (i) the value of a model can only be determined in a defined context, and that (ii) even though no model is 100% correct, its use may be justified in some situations.

The classical elements of toxicological validation are not sufficient to describe the usefulness of animals as models. Although robustness, predictivity and relevance are important, they need to be combined with sharp definitions of the model purpose and its applicability domain. Even with all this done, some further aspects need to be considered.

First, a model is always in competition with other models. How good are these? The answer determines the "necessity" and the "competitive advantage" of a model.

Second, no model only has advantages, it always also has costs and disadvantages. This information needs to be factored in to understand the overall "benefit". Even a broken clock is right twice a day. This is useful at these time points, but it can be confusing at others. (Fig. 1)

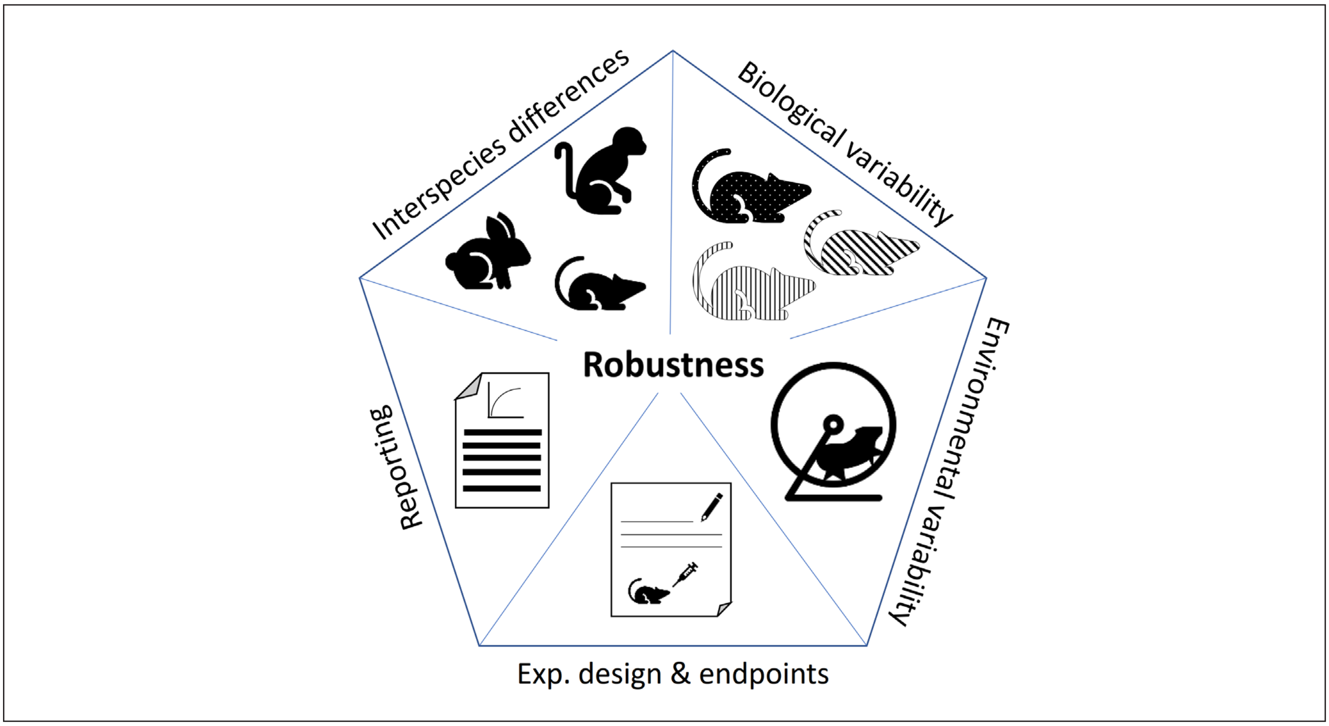

Fig. 3: Exemplification of aspects that determine the robustness of animal-based test methods
Five important categories are shown: 1) biological variability of the animals used (and their hygienic/health status), 2) variability caused by fluctuations of environmental parameters (e.g., food, noise, housing), 3) variability arising from the experimental design (e.g., statistics, risk of bias), the experimental procedures, and the quantification of test endpoints, 4) variability arising at the stage of data processing (e.g., dealing with outliers), data presentation, and post-hoc analyses, and 5) uncertainty arising from the choice of the relevant animal model (strain, species, etc.).

Third, “measurement units” need to be established. Is “usefulness” measured in monetary value, in some form of ethical currency, in economic terms, in measures of reputation and fame, etc.?

Fourth, once the measurement unit has been determined, the way to assess this needs to be defined. This item is less trivial than it sounds. Often, the question is: How good is good enough? Is a predictivity of 50% good? What about 5%? Is it sufficient to be qualitatively correct, or is a quantitative statement required?

In this process, it also needs to be decided how much weight should be given to different parameters and how quantitative the measurement methods should be. Very important is also whether one considers the immediate future, the distant future, or a combination thereof. In the current discussion, anecdotal evidence (single cases from the past) is often used as a major line of argument. A scientific approach could be an alternative: A quantification of benefit and necessity may be performed for many fields. Such data may be combined in a meta-analysis using methods established for systematic reviews. Such a comprehensive analysis is not yet available, but it is being attempted for some fields, and a committee has been set up by the National Academy of Sciences of the USA to, e.g., judge the robustness of animal models.[1]

3 The knockout criterion: robustness

Model requirements in general

A model can be characterized in many ways, and what often comes to mind first is the performance of a model (specificity, sensitivity, accuracy, etc. of predictions), i.e., its predictivity. The internal construction of the model, i.e., its relevance, is only considered as second thought: How does it work, which aspects of the reality does it reflect well or less well, are the outcomes observed in the model and in reality based on the same internal steps and functional relationships of its elements, etc.? Robustness, the most important criterion, the “*one to rule them all*” (to cite a passage from Tolkien’s *Lord of the Rings*) is often neglected, as it is taken for granted. What is this fundamental criterion on which all others are built – the *conditio sine qua non*? The importance of robustness has been stressed in the editorial to an entire journal volume on animal models and their validity (Pound, 2020).

If a model is robust, it will produce the same output/response when given the “same” input (Leist et al., 2008, 2014). This means that the results should be independent of the experimenter

[1] https://bit.ly/3JO3aBv

using the model, or the time, or the place. If the model is the basis of a test method, its robustness is described in terms of variance between technical replicates, between independent runs within a laboratory, and as interlaboratory reproducibility.

Factors affecting robustness

The above examples indicate that "same" input conditions are hard to define. In case of animal models, several animal-specific parameters need to be considered that can influence the output of a model: the strain, the provider of the strain, the feed, the caging, the bedding, etc. (Fig. 3). If one thinks further, the various non-animal parameters such as environmental lighting, sound, smell, temperature, experimental dosing, solvent, handling, etc. already indicate that replicating exactly the "same" input is impossible.

Variability of input parameters of animal models has two fundamentally different reasons: First, it is not possible to completely avoid all variation in parameters known to potentially affect the model and the outcome of an experiment. Second, there are always unknown parameters that may affect an experiment and that are not controlled. Examples for the latter point include, e.g., scents of the cosmetic and personal care products used by the animal caretaker, a fight between animals having happened before the experiment, a small failure/mistake in the supply chain/production procedure that generated the feed or bedding, etc.

Since not all parameters can be controlled, "robustness" not only measure the highly reproducible functioning of the model, but also implies some resilience to unavoidable variations of conditions.

The influence of uncertainty on robustness

Some parameter variations are unavoidable because of the stochastic nature of all physical interactions, chemical reactions, and biological phenomena. This condition means that each output value of an experiment is a distribution function of values that is very narrow around an average in optimal cases, or broad in many other cases. The uncertainty arising from this is called aleatoric uncertainty (from *alea* = the dice). On top of this, variability may be further increased by a lack of knowledge (epistemic uncertainty). For instance, in a litter of mice, it may make a difference in which position a given pup was in the uterus, at which time it was delivered relative to its litter mates, how its access was to mothers' milk, etc. We do not usually have this type of information, but such parameters contribute to a size and age variation within a litter of genetically identical mice. This may, in the end, affect the variability of outcomes in animal experiments.

How relevant are such theoretical considerations in the real world? Is there an issue with the robustness of animal models? Are there data on the variability of different parameters? The issue is complex, as measures of variability need to be agreed upon, and they will strongly depend on the endpoints assessed and the model predictions required. However, it is broadly recognized in the field that there is such an issue (Hartung and Leist, 2008; Busquet et al., 2020; Daneshian et al., 2015). For instance, CAMARADES centres[2] address this point, and it has been suggested that the classical 3Rs principle, replacement, reduction and refinement, should be expanded by three further Rs (robustness, registration and reporting) for animal studies (Strech and Dirnagl, 2019). The fact that large initiatives deal with the issue suggests that the robustness of animal studies can be a real-world problem.

However, there is also evidence that this may not necessarily be an inherent problem of high uncertainty of data derived from animals as a model (Kafkafi et al., 2018). It has rather been found that the robustness of research data is affected by aspects other than the aleatory uncertainty (variations in the system due to insufficient stability). There are also "human" factors, including weaknesses in experimental design, incomplete data reporting, and inadequate dealing with the risk of bias. Such shortcomings may be avoided by implementing adequate rules for experimenters and their institutions, including mandatory study preregistration (Begley et al., 2015; Vogt et al., 2016; Bert et al., 2019).

There is a broad scientific debate on the potential to reduce the variability in animal studies by standardization (Richter et al., 2009, 2010). A critical outcome of respective studies is that reducing variance by rigid standardization may negatively impact the predictivity of animal models (Voelkl et al., 2021). This may require more consideration in the experimental design (Voelkl et al., 2020).

It has also been noted that poor reporting affects the robustness of animal models (Kilkenny et al., 2010). This problem should not occur in studies complying with the updated ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines (Percie du Sert et al., 2020).

Robustness vs predictivity

In the discussion on robustness, the term as such has not only various aspects (see above) in its classical sense, but it is sometimes also used in the sense of "reliability of prediction" (van der Worp et al., 2010). The latter is strongly related to another model feature, predictivity. This fuzzy use of concepts in the literature shows that the various parameters of a model are in fact not fully independent of one another and are sometimes hard to separate (predictivity cannot be defined in the absence of robustness, while robustness can only be measured when some prediction goal is considered).

Examples of model robustness

Toxicology is an area that is particularly suitable to define robustness of animal models as study protocols are highly standardized, several data sets on repeated experiments are available (e.g., toxicity of chemical X, measured by companies Y and Z), and rules on documentation have been established under the framework of good laboratory practice (GLP). Moreover, there is less uncertainty in looking at a potential adverse outcome (2-dimensional uncertainty: animal physiology x drug effects), than

[2] https://www.ed.ac.uk/clinical-brain-sciences/research/camarades/about-camarades

in looking at disease modification by a drug in pharmacological studies (3-dimensional uncertainty: animal physiology x disease model x drug action).

Some toxicological data sets suggest that the robustness of animal models can be less than ideal. For instance, data collected on the uterotrophic assay in the context of the US Environmental Protection Agency (EPA) endocrine disruptor program showed a very high variability of animal data (Browne et al., 2015). Similar observations were made, e.g., for the Draize eye irritation test data in the European Chemicals Agency (ECHA) database (Luechtefeld et al., 2016a), where in substances tested multiple times in the same animal model there was a 10% chance of a non-irritant evaluation after a prior severe-irritant result according to UN GHS classification criteria. In contrast, some areas of toxicology show that animal studies can yield robust data, such as for oral toxicity data in the same database (Luechtefeld et al., 2016b). However, it is important to note that few studies quantify the precision of animal models so that one can rationally discuss what "robust" means for this field. Another important approach was the mining of ToxRefDB (one of the world's largest *in vivo* toxicology databases curated and run by the US EPA) for studies (> 3500 chemicals) on repeated-dose toxicity (Ly Pham et al., 2020). It found that animal data correlate with one another with a maximal r^2 of 0.55-0.7. This finding would mean that a toxicity threshold of 10 mg/kg/day found in one study may in reality mean anything from 1-100 mg/kg/day (95% confidence interval).

More data on robustness may have been collected by pharmaceutical companies that have used animal models for drug screening. The discovery of many psychiatric drugs and their successful use avoiding cardiotoxicity (long QT syndrome, torsades de points) suggests good robustness in such areas.

In summary, the background data on robustness of animal models are relatively scarce, and they possibly differ strongly between application areas. This aspect suggests that follow-up studies are needed to clarify how robust different types of animal models are.

Inter-species robustness

Toxicological studies can also address the important question how robust data from animal studies are when species or strains are compared. Some limited but dramatic examples show that there are such species differences, e.g., of mice vs New Zealand white rabbits concerning the toxicology of thalidomide or of different mouse strains concerning hepatotoxicity of arylhydrocarbon (Ah) receptor ligands. Leaving examples that can be easily explained by known physiological differences (transporters, metabolizing enzymes, etc.) aside, there are still issues of higher susceptibilities of, e.g., rats over mice for chemical-induced peripheral neuropathies or for mice over rats concerning liver carcinogenesis. In general, large toxicological areas (like developmental and reproductive toxicity) show inter-species robustness that is as low as 60% (Smirnova et al., 2018). Large variation has also been observed in the bioavailability of chemicals (Grass and Sinko, 2020) or in chronic toxicity endpoints recorded during cancer studies of mice and rats (Wang and Gray, 2015). Although this is discussed here mainly as a robustness issue (variability between animals), such data are also important for the predictivity discussion: In cases where there are pronounced differences between species (or strains): Which one should be chosen for prediction of human effects?

4 Conclusion and outlook

It is important to note that we are dealing here mainly with the structuring of arguments, not with the data themselves. Moreover, our discussion does not address research on animals as such, as in veterinary and basic biology research projects, but only on research where animals are taken as models of humans.

In this first part of the article mini-series, we have discussed robustness as one of the main parameters that needs to be considered for an unbiased and comprehensive evaluation of animal model usefulness. The aspects related to the benefit and necessity of a particular animal model (Fig. 1), based on the chosen metrics and evaluation strategy (Fig. 2), will be addressed in a later BenchMarks article.

Eventually, an approach supported by solid data and a transparent process on how to integrate them will be required for a realistic evaluation of the usefulness of animal-based tests. However, it should already have become clear that animal models should be evaluated like any other model. This is often still not the case. Animals are considered the gold standard in several research fields; and even today, reification[3] of animal models is widespread in scientific and political discussions (sometimes by mistake/misconception; sometimes on purpose). A recent detailed report by the UK's All-Party Parliamentary Group (APPG) for human-relevant science made clear that misconceptions on the usefulness of animal experiments may be one of the fundamental reasons preventing medical progress through the use of other, non-animal models.[4]

References

Begley, C., Buchan, A. and Dirnagl, U. (2015). Robust research: Institutions must do their part for reproducibility. *Nature 525*, 25-27. doi:10.1038/525025a

Bert, B., Heinl, C., Chmielewska, J. et al. (2019). Refining animal research: The animal study registry. *PLoS Biol 17*, e3000463. doi:10.1371/journal.pbio.3000463

Browne, P., Judson, R. S., Casey, W. M. et al. (2015). Screening chemicals for estrogen receptor bioactivity using a computa-

[3] Reification means that model and reality (or idea vs real thing) are obfuscated. A common case of reification is the confusion of a model with reality, paraphrasing Alfred Korzybski, "*the map is not the territory, and an Alzheimer mouse is not an Alzheimer patient*".

[4] https://bit.ly/3wNx3yw

tional model. *Environ Sci Technol 49*, 8804-8814. doi:10.1021/acs.est.5b02641

Busquet, F., Kleensang, A., Rovida, C. et al. (2020). New European Union statistics on laboratory animal use – What really counts! *ALTEX 37*, 167-186. doi:10.14573/altex.2003241

Daneshian, M., Busquet, F., Hartung, T. et al. (2015). Animal use for science in Europe. *ALTEX 32*, 261-274. doi:10.14573/altex.1509081

Ferreira, G. S., Veening-Griffioen, D. H., Boon, W. P. C. et al. (2020). Levelling the translational gap for animal to human efficacy data. *Animals 10*, 1199. doi:10.3390/ani10071199

Grass, G. M. and Sinko, P. J. (2020). Physiologically-based pharmacokinetic simulation modelling. *Adv Drug Deliv Rev 54*, 433-451. doi:10.1016/s0169-409x(02)00013-3

Hartung, T. (2007). Food for thought ... on validation. *ALTEX 24*, 67-80. doi:10.14573/altex.2007.2.67

Hartung, T. and Leist, M. (2008). Food for thought ... on the evolution of toxicology and the phasing out of animal testing. *ALTEX 25*, 91-102. doi:10.14573/altex.2008.2.91

Hartung, T. (2010). Evidence-based toxicology – The toolbox of validation for the 21st century? *ALTEX 27*, 253-263. doi:10.14573/altex.2010.4.253

Hartung, T., Hoffmann, S. and Stephens, M. (2013). Mechanistic validation. *ALTEX 30*, 119-130. doi:10.14573/altex.2013.2.119

Judson, R., Kavlock, R., Martin, M. et al. (2013). Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. *ALTEX 30*, 51-56. doi:10.14573/altex.2013.1.051

Kafkafi, N., Agassi, J., Chesler, E. J. et al. (2018). Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neurosci Biobehav Rev 87*, 218-232. doi:10.1016/j.neubiorev.2018.01.003

Kilkenny, C., Browne, W. J., Cuthill, I. C. et al. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol 8*, e1000412. doi:10.1371/journal.pbio.1000412

Krebs, A., van Vugt-Lussenburg, B. M. A., Waldmann, T. et al. (2020). The EU-ToxRisk method documentation, data processing and chemical testing pipeline for the regulatory use of new approach methods. *Arch Toxicol 94*, 2435-2461. doi:10.1007/s00204-020-02802-6

Leist, M., Hartung, T. and Nicotera, P. (2008). The dawning of a new age of toxicology. *ALTEX 25*, 103-114. doi:10.14573/altex.2008.2.103

Leist, M., Hasiwa, N., Rovida, C. et al. (2014). Consensus report on the future of animal-free systemic toxicity testing. *ALTEX 31*, 341-356. doi:10.14573/altex.1406091

Luechtefeld, T., Maertens, A., Russo, D. P. et al. (2016a). Analysis of Draize eye irritation testing and its prediction by mining publicly available 2008-2014 REACH data. *ALTEX 33*, 123-134. doi:10.14573/altex.1510053

Luechtefeld, T., Maertens, A., Russo, D. P. et al. (2016b). Analysis of public oral toxicity data from REACH registrations 2008-2014. *ALTEX 33*, 111-122. doi:10.14573/altex.1510054

Ly Pham, L., Watford, S., Pradeep, P. et al. (2020). Variability in in vivo studies: Defining the upper limit of performance for predictions of systemic effect levels. *Comput Toxicol 15*, 1-100126. doi:10.1016/j.comtox.2020.100126

Percie du Sert, N., Hurst, V., Ahluwalia, A. et al. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *BMJ Open Sci 4*, e100115. doi:10.1136/bmjos-2020-100115

Pistollato, F., Bernasconi, C., McCarthy, J. et al. (2020). Alzheimer's disease, and breast and prostate cancer research: Translational failures and the importance to monitor outputs and impact of funded research. *Animals 10*, 1194. doi:10.3390/ani10071194

Pound, P. (2020). Are animal models needed to discover, develop and test pharmaceutical drugs for humans in the 21st century? *Animals 10*, 2455. doi:10.3390/ani10122455

Richter, S. H., Garner, J. P. and Würbel, H. (2009). Environmental standardization: Cure or cause of poor reproducibility in animal experiments? *Nat Methods 6*, 257-261. doi:10.1038/nmeth.1312

Richter, S. H., Garner, J. P., Auer, C. et al. (2010). Systematic variation improves reproducibility of animal experiments. *Nat Methods 7*, 167-168. doi:10.1038/nmeth0310-167

Smirnova, L., Kleinstreuer, N., Corvi, R. et al. (2018). 3S – Systematic, systemic, and systems biology and toxicology. *ALTEX 35*, 139-162. doi:10.14573/altex.1804051

Strech, D. and Dirnagl, U. (2019). 3Rs missing: Animal research without scientific value is unethical. *BMJ Open Sci 3*, e000035. doi:10.1136/ bmjos-2018-000048

van der Worp, H. B., Howells, D. W., Sena, E. S. et al. (2010). Can animal models of disease reliably inform human studies? *PLoS Med 7*, e1000245. doi:10.1371/journal.pmed.1000245

Voelkl, B., Altman, N. S., Forsman, A. et al. (2020). Reproducibility of animal research in light of biological variation. *Nat Rev Neurosci 21*, 384-393. doi:10.1038/s41583-020-0313-3

Voelkl, B., Würbel, H., Krzywinski, M. et al. (2021). The standardization fallacy. *Nat Methods 18*, 5-7 doi:10.1038/s41592-020-01036-9

Vogt, L., Reichlin, T. S., Nathues, C. et al. (2016). Authorization of animal experiments is based on confidence rather than evidence of scientific rigor. *PLoS Biol 14*, e2000598. doi:10.1371/journal.pbio.2000598

Wang, B. and Gray, G. (2015). Concordance of noncarcinogenic endpoints in rodent chemical bioassays. *Risk Anal 35*, 1154-1166. doi:10.1111/risa.12314

Acknowledgements

This work was supported by CEFIC, the DFG, the BMBF, the Land-BW (NAM-ACCEPT), and funding by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 964537 (RISK-HUNT*3R*), No. 964518 (ToxFree) and No. 825759 (ENDpoiNTs).