Bench Marks!

On the Usefulness of Animals as a Model System (Part II): Considering Benefits within Distinct Use Domains

Giorgia Pallocca[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

In many countries, animal experiments can only be performed when their necessity has been demonstrated in a legal document. As the usefulness of animals in research is also a significant societal and political issue, criteria to structure debates and evaluations are needed. Here, background information is given on laboratory animal studies. Moreover, parameters that may be considered in judging their usefulness are suggested. The discussion is strictly focused on animals used as tools/test systems/models to provide information on humans. In this context, general features and performance characteristics of models are discussed. Examples are given for well-recognized criteria (e.g., robustness, relevance, predictivity) to judge the usefulness of predictive models. The main hypothesis put forward here is that a benefits evaluation (usefulness metrics) is only possible within sharply circumscribed "use domains". Examples are given for the research fields of drug and vaccine research, toxicology, disease pathogenesis, and basic biological research. Efficacy, safety, and quality studies are highlighted as "use domains" within the field of drug discovery and production. A further separation into individual diseases, drug targets or symptoms is suggested for, e.g., efficacy studies or pathophysiology. Finally, an outlook is given on the evaluation of model advantages and disadvantages to arrive at their "net benefit". Moreover, the need to compare the net benefits of animal models versus that of their alternatives is highlighted.

1 Introduction

Usefulness sounds like an easy concept. Thus, the "usefulness of animals" should be easy to determine. After some consideration, it becomes clear that the general question of usefulness is so complex and undefined that there is no reasonable approach to define it and no metrics to judge it. The question of usefulness should always be followed by a definition of the *purpose of use*. This description should be as sharp as possible. Once the *use domain* is known, the question becomes more defined and answerable. Follow-up questions would be: Who defines usefulness? And what parameters are used to judge usefulness?

To make this less theoretical, we introduce examples from the everyday world (outside science, considering well-known leisure activities). They may help to illustrate a scientific problem, and this way we hope to offer an easy entry into a complex discussion. For instance, one may ask whether skis are a useful tool. It becomes immediately evident that this question is pointless without specification of a purpose (e.g., opening a bottle, descending a mountain slope, crossing the jungle). Even when the purpose (ascending or descending a mountain slope) is defined, the tool's usefulness is determined by the exact situation (snow coverage, etc.). Another easily accessible example is the use of a full-face helmet. It may be argued that helmets are very useful because

Received July 11, 2022;

ALTEX 39(3), 531-539. doi:10.14573/altex.2207111

Correspondence: Marcel Leist, Ph.D.
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)

they offer protection. Nevertheless, most people reading this article in their office will not be wearing a helmet. They may put on their helmet when going out and riding their motorbike, and they may use some type of helmet when riding a horse or a bicycle, but not a full-face helmet. The reason is that the net benefit (balance of advantages and disadvantages) of a very protective full-face helmet is not high in the latter examples (Fig. 1).

What is the point of introducing these trivialities at the beginning of a scientific article? They help to show that the question on the *usefulness of animals* is more complex than it appears at first sight. In this situation, we consider it helpful to approach it at different levels. The examples promote an intuitive understanding (not requiring abstract thinking). We hope that such an approach to the question will increase the motivation to invest time and effort to address the main topic of this article.

In this context, it is important to note that some of the problems already have been outlined in part I of this article series (Pallocca et al., 2022a). Altogether, six elements of the question will be covered in this series (i) consistency of animal-derived data (robustness of the model system; part I); (ii) scientific domain investigated (e.g., toxicology vs disease modelling vs therapy; part II, here); (iii) "net benefit" (integrating positive and negative aspects; part III); (iv) benchmarking to alternatives; (v) ideas for usefulness metrics (How good is good enough?); (vi) procedures to assess benefit and necessity (Fig. 2).

We would like to reiterate that we are not discussing the general *usefulness of animals* here, but, more specifically, the "usefulness of animals as models" (Fig. 3). The difference is clear to many specialists in the area who work with animal models or use alternative models for animals. However, practical experience from dozens of public discussions shows that various stakeholder groups still have many misconceptions. Reification of animal models is the most severe amongst these. "Reification" is the technical term for confusing a model with reality. A frequently cited reification example is taking a map for being the landscape. A subtler but common example is to confuse a clinical drug trial on 20 healthy young volunteers with the situation of a large country's population being treated with a drug on prescription by general practitioners.

In the field of animal models, both weak and strong reification are widespread. "Strong reification" refers to the assumption that an animal behaves exactly like the human population and that a reaction in an animal model is the same as in human disease. Although this concept is severely flawed, it is still widespread in some disciplines. One explanation may be that proponents of this opinion are often unaware of their mistakes. The assumption is simply part of some cultures, and therefore it is passed on without being reconsidered or challenged (like the medieval belief that serious diseases are cured by bloodletting). A "weak reification" refers to the assumption that animals are indeed only a model for humans but that either (i) they need to be considered as the gold standard to calibrate other models or (ii) that data derived from animals may not be directly transferable to humans, but that it is better to have animal-derived information than nothing. From a practical point of view, this attitude can make sense. Often, animal data are indeed the only

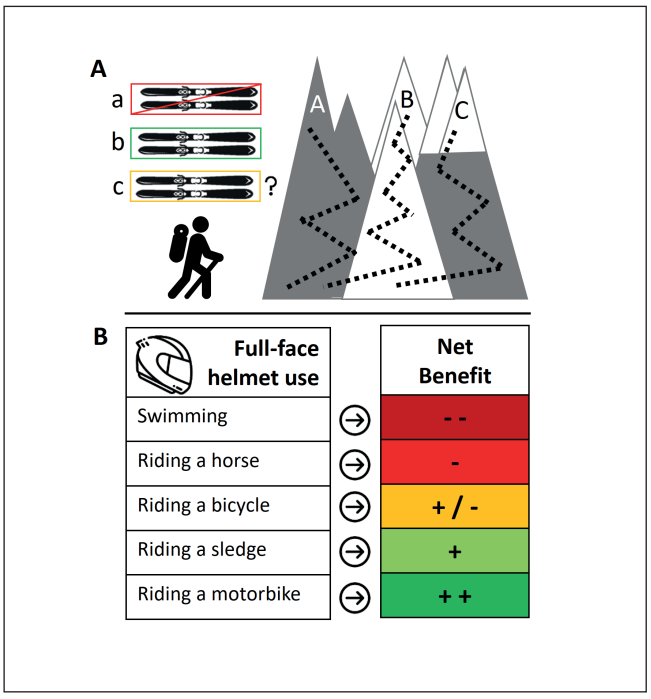

Fig. 1: Exemplification of the net benefit of a "tool" depending on the use domain

A. Three mountain paths are shown (A, B, C). A is not covered in snow, B is entirely snow-covered, and C is partially snow-covered. The tool "ski" has benefits (quick progression) on snow and disadvantages (heavy to carry) when there is no snow. In situation A, there is no net benefit, and the tool would not be useful (a). In situation B, there is a net benefit, and the tool would be useful (b). In situation C, the tool may be useful, but more information is required on the situation and the weight of advantages and disadvantages (c). B. A full-face helmet has advantages (protection) and disadvantages (weight, reduced view). Depending on the situation, it has a net benefit (+, ++) or a net disadvantage (-, --).

data available. This aspect can make them important and precious, even though it is not known which data can be transferred to humans and which not. Careful use of such data as the gold standard can help to advance the technical development of other models. It can also give new ideas on disease targets or pathological mechanisms. And it may then be tested in other models to determine whether these are relevant. However, there is also a downside. If the approach is used too strictly, new models that predict humans better than animal models may fail to be accepted because they do not predict animal outcomes. This problem is well known for toxicological assays. Examples can also be found in disease research and drug discovery.

The take-home message from the above discussion is that animal usefulness is considered here in the context of its model character. What is meant by "model character"? Or, to reverse the perspective, what would be an aspect that does not have a model character? An example may be helpful to prepare for the answer. Let's consider a canary bird. The animal may be regarded as a

tool (to detect methane in coal mines), as a model (to study pesticide ecotoxicity), as a pet (that feels lonely), as a breeding object (that wins a prize), as a study object (with specific song behavior), as part of an ecosystem (e.g., prey for others), etc. (Fig. 3). This example shows that one word/thing can have many connotations and use aspects. Here, we only discuss the function of animals as tools/models/test systems. Thus, we do not deal with animals in scientific fields that address animal physiology, evolution, behavior, and social context as such.

This perspective has direct implications on the judgement of usefulness. All discussion aspects must relate to the model character of animals, and they need to use the fundamental and generally agreed criteria established for model evaluation in dozens of other, otherwise unrelated fields. All models must be robust, predictive, and relevant to some degree. The robustness/reliability aspect has been discussed previously (Pallocca et al., 2022a) as the *conditio sine qua non*. Without robustness of a model system, other parameters cannot be evaluated (for technical reasons) and should not be assessed (for efficiency reasons) (Fig. 2). Being such a fundamental feature, robustness is sometimes also termed "internal validity".

The metrics of how precisely results from the model can be transferred to a real-life situation (i.e., to clinical drug effects, to human pathophysiology, or other human-relevant aspects) is called predictivity. To obtain this measure, data are compared to the world outside the model, and therefore predictivity is sometimes called "external validity".

The third feature of test methods is relevance. The concept refers to the internal working of a model, and it is particularly hard to assess and quantify. In the future, we expect that more research will focus on the relevance of models, not just regarding animals but also to a large degree in the field of new approach methodology (NAM). In the past, this consideration has been neglected, and, for lack of better tools, "predictivity" has often been used as a proxy to quantify "relevance"[1].

In summary, we intended to introduce the following concepts: (i) Animals are considered here only in their function as models to predict humans or as tools to promote human-related research (understanding human physiology, pharmacology, behavior, toxicology, etc.). (ii) It is essential to understand the model/tool character and the performance criteria linked to this concept. (iii) Within the perspective of animals as models, it is pivotal to look at usefulness in sharply-defined "use domains" (as a general usefulness definition gets blurry and escapes all attempts to define practically-useful evaluation criteria). (iv) When using criteria of benefit, it is important to factor in advantages and disadvantages to arrive at a "net benefit". (v) Last but not least, this overview is written from the perspective of non-users of animals. Is this meaningful? We think that it can provide a fresh view and new arguments. There are countless articles written by animal users on animal models (and also by animal non-users about animal-free methods). We feel that more cross-interactions are required to come to a balanced discussion. We presume that all

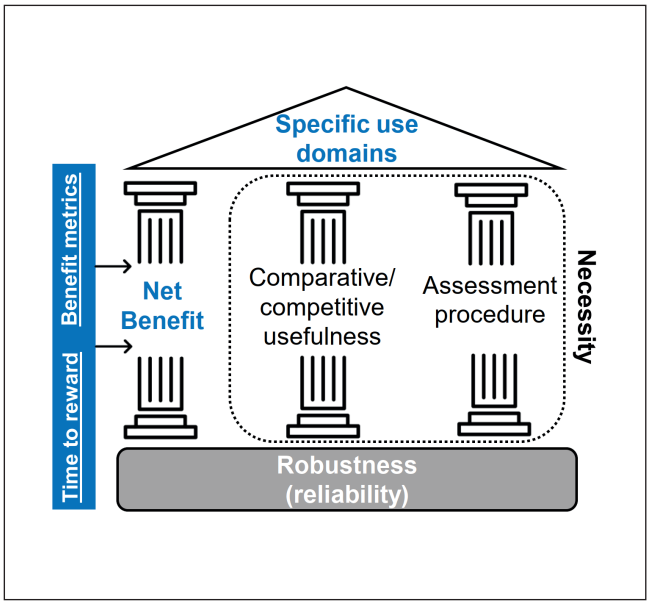

Fig. 2: Critical aspects of the usefulness assessment of a test method

The foundation for the usefulness is reliability (also called robustness, technical reproducibility, or internal validity). This aspect has been discussed with respect to animals as models in part I of this article series (Pallocca et al., 2022a). The overall umbrella (roof of the construction) is the specific use domain. A method can only be judged (on all aspects) concerning its purpose and applicability domain. If a method has several use domains, it needs to be considered for each of them separately. This aspect is discussed here in part II of this series. Specific examples from a panel of use domains are encouraged as contributions from the broad community in an extension of this article series. One of the central pillars is the net benefit, i.e., the sum of advantages and disadvantages of the method. A judgement on this requires some benefit metrics, which need to consider the predictivity and relevance of the method for the defined purpose (use domain). What makes the quantification of net benefit challenging is the time horizon, i.e., which time frame is used to score benefits and disadvantages. Some may be immediate, or they might be envisaged for the near future. Others may have a long lag period. The overall score may differ depending on the time point (or period) of the assessment. Such aspects will be discussed in part III of the series. A second important pillar is the comparative usefulness of a method. This refers to the fact that there may often be alternative or competitive methods, and in real life it is not the absolute usefulness that counts but the usefulness in comparison to alternatives (i.e., the "necessity" to use the particular method). The third pillar refers to the assessment procedure, i.e., the type of metrics chosen, the way to retrieve the information, the groups contributing to the decision process, and the way the decision process is organized. The second and third pillars will be discussed later in this series.

[1] In this sentence, the key words were accidentally swapped in part I (Pallocca et al., 2022a).

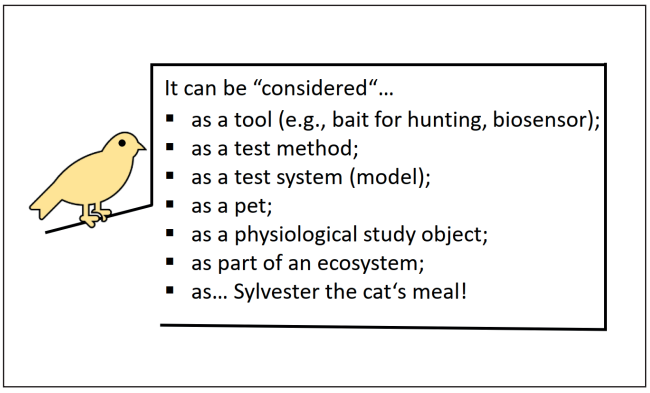

Fig. 3: Exemplification of "animal models" as one of the concepts of animal use

Here a canary (bird) is used as an example that a single word can have various connotations, i.e., its meaning can be considered from different perspectives. It needs an agreement between participants of a discussion on the perspective to be considered, compared, and evaluated. For instance, a canary can be considered as a pet. It may also be considered a study object for learning about canary physiology, behavior, or perhaps canary evolution. Others may consider it in its role in a given ecosystem and for interactions with other elements in this system. Such perspectives are not considered here in the usefulness evaluation. They are different from, e.g., considering a canary as a tool (e.g., to monitor methane concentrations in a coal mine). There are differences between use as a *model*, *tool*, *or test method*. However, these three concepts/perspectives also overlap and are therefore sometimes used interchangeably. The same would apply to using *animals as a model*, *tool*, *test system*, *or test method*. They all differ from the view of a pet or direct study object in the sense that animals are used to achieve information on something outside the animal. The use of animals (or a canary) as part of a test method (as the test system or model system of the method) is the primary role of laboratory animals. This is not only the case in toxicology and drug development. Also in basic biological research, researchers using mice are not ultimately interested in, e.g., mouse aging or mouse microbiomes; they use the mouse as a model to conclude on humans. Exceptions exist for which the usefulness evaluation suggested here does not apply strictly.

stakeholders may reach an agreement that animals may be valuable in one domain but not in another. In contrast, agreement on a limitless acceptance or a total ban of animals is unlikely. We hope that different sides would more easily come to a consensus if advantages and disadvantages were compared within given use domains and if some criteria were agreed upon on how to judge and weigh the arguments. In the end, this exchange could benefit the field of animal experimentation and help the non-animal NAM developers to focus efforts and avoid mistakes known from animal experimentation. Model performance characteristics must (of course) also be considered for NAM, and the problem of reification also exists in this field.

2 Use domains: different dimensions of usefulness

2.1 An appeal to be specific about the purpose of animal use

The most significant source of misunderstandings and confusion in public debates relates to the purpose of a model. It appears trivial that the usefulness of a model or a test method can only be judged if its objective is known. It should thus be common practice to use the purpose (background questions) as the main guidance during the evaluation of an animal model. Nevertheless, it can be frequently observed in discussions that this principle is violated or neglected. We would like to recall here that any test method (the term "test method" is used here synonymously with model for simplification reasons) is defined by an exposure scheme, a test system, a test endpoint, and a prediction model (sometimes called data interpretation procedure) (Leist et al., 2010; Schmidt et al., 2017; Krebs et al., 2020; Bal-Price et al., 2018; Worth and Balls, 2001; Griesinger et al., 2016). What is often forgotten is that this only makes sense when the underlying question to be solved is defined, i.e., the use of the test method is explicit (Krebs et al., 2020). When the use domain's anchoring is lost, the arguments in favor of the model do not refer to the same use(s) as those that deny the usefulness of animals. In other words, the principle of balancing advantages and disadvantages cannot be applied. In the classical analogy of a beam balance, the beam represents a given use domain. The balance does not work if advantages belong to one beam and disadvantages to another (Fig. 5). This is like "comparing apples and oranges", and it is, unfortunately, rare to observe a debate where such mistakes are not made. In about two dozen interviews with journalists, we have rarely experienced a clear understanding of this issue/fallacy.

Typical mistakes in this field are comparing failures of animal experiments (e.g., in predicting the reproductive toxicity of thalidomide) with the benefits/usefulness of animal experiments in discovering new drugs to treat, e.g., the symptoms of Parkinson's disease. The first question refers to a toxicological purpose, while the second argument's objective is to define the efficacy of drug candidates. These are different use domains. In theory, it would be possible to work entirely without animals on drug efficacy but to use animals for drug safety – or the other way round. Although the efficacy and safety domains look pretty similar from the outside, they have different objectives, legal frameworks, and working cultures. They should therefore be judged separately. Cost-benefit analysis across such different domains is a challenge that has not been solved yet. For this practical reason, the usefulness of animal experiments can and should only be discussed and evaluated *within clearly defined purpose domains*. The broader these domains get, the more difficult it is to judge "in general" their robustness, reliability, relevance, predictivity, and the overall ratio of "costs" and "benefits". Conversely, it is realistic that the critical performance parameters of animal models (robustness and net benefit) can be judged within sharply defined use domains. In addition, focusing on narrow, clearly delineated domains allows an evaluation of how necessary animal studies are in these fields, i.e., how they perform relative to alternative approaches available for the respective use area (Fig. 2, 4)

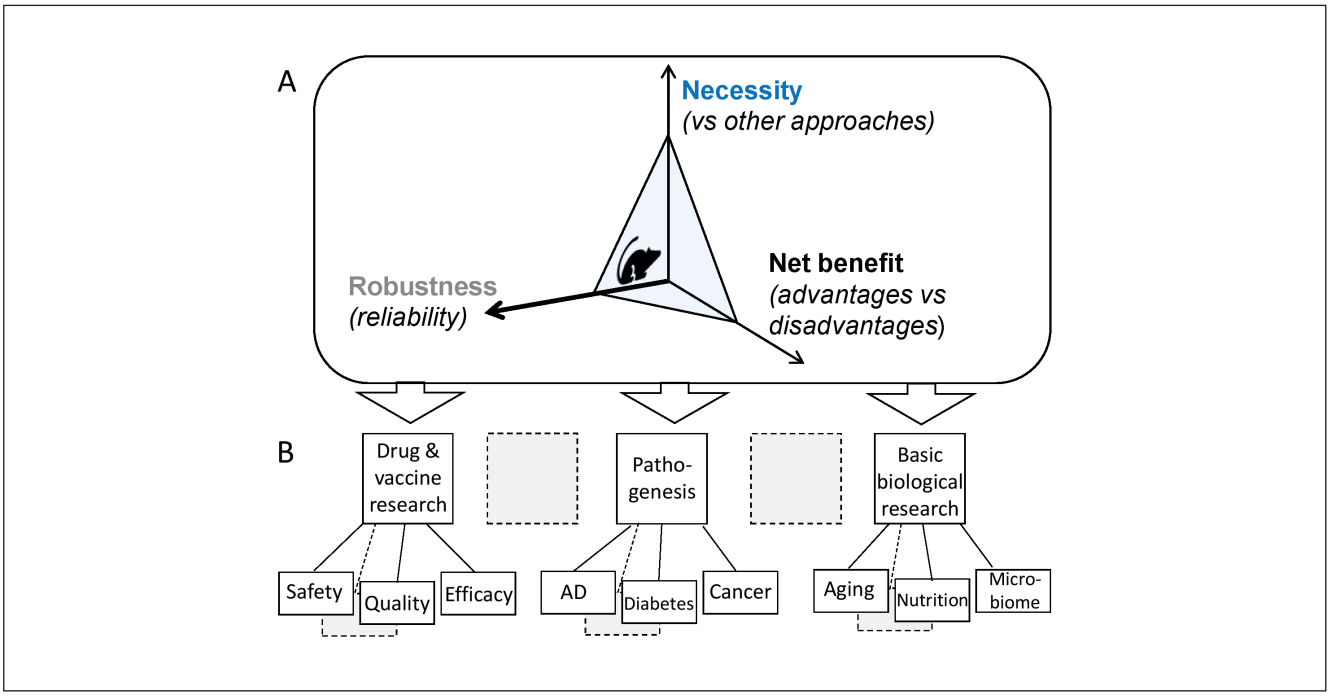

Fig. 4: Schematic overview of usefulness evaluation for various use domains

A. Major dimensions of the usefulness evaluation are shown: robustness, net benefit and necessity (= competitive advantage vs other methods). These are considered for each use domain. B. Exemplification of 12 use domains (out of many more): Use domains may be grouped into larger units (5 examples are shown). For instance, research into the efficacy, quality, or safety of drugs may be grouped as drug research, or research concerning aging, nutrition, and the role of the microbiome may be grouped as biological research. The usefulness of animals as a model or test method can only be judged for smaller subdomains. For instance, it may be found that animals are useful for cancer research in the group on pathogenesis, but they are not useful for research in Alzheimer's disease (AD). Sometimes even smaller divisions may be necessary. For instance, animals may be useful for finding symptomatic treatments for epilepsy or Parkinson's disease, but less useful for treatments that cure the diseases.

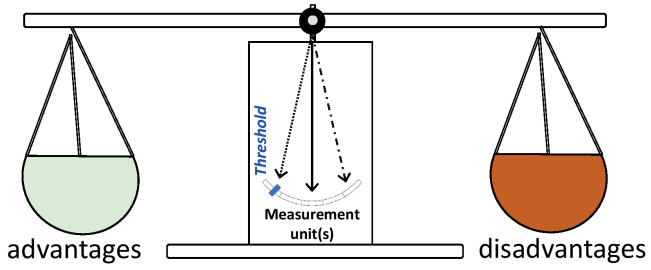

Fig. 5: Schematic representation of net benefit

The construct of balancing advantages/benefits and disadvantages/shortcomings should ideally be applied to each narrowly circumscribed use domain (understanding of a certain disease, safety of a group of drugs, etc.) to define its overall usefulness. The whole process of weighing can only be performed if there is a solid enough basis, i.e., robustness of the model/method as such. The process can only be performed if the advantages and disadvantages are from within the use domain. This means that the balance beam represents only one use domain. If there are many use domains, the balancing act must be repeated within each one. The process will, in most cases, be qualitative or semi-quantitative. But even if no full quantification is desired, there needs to be some consensus on the measurement unit(s) applied. This may comprise ethical factors, economic considerations, scientific/technological progress, or aspects of individual freedom. Moreover, threshold criteria would need to be defined, i.e., how much balance is necessary and how much misbalance is acceptable.

2.2 Animals in drug and vaccine research

One major area of animal use is the discovery, development, and production of drugs. This domain is named here first, as it receives the most public attention. Notably, the field is large and very heterogeneous. It would not even be clearly circumscribed if only a single disease was considered. Animals are used for widely different purposes in the drug field to address efficacy, safety, and quality questions (Busquet et al., 2020a). The first two areas are considered for each drug or each disease area, the last mainly during drug production (Fig. 4).

As the three use areas are often confused, a short explanation is given here, followed by an explanation of why these areas are to a large extent fully independent of one another. *Efficacy* deals with the desired activity of drugs, i.e., it asks whether drugs affect disease symptoms or progression or have another (e.g., palliative) positive effect. *Safety/toxicity* deals with unwanted side effects of drugs (e.g., a diabetes drug giving skin rashes, causing nausea, promoting cancer, or triggering fetal malformations). Sometimes, the side effects are linked to the drug efficacy (e.g., a diabetes drug reducing blood sugar so strongly that patients feel nauseous), but this will not be discussed here for simplicity reasons. *Quality* deals neither with safety nor efficacy but with the drug production process. Relevant aspects include the purity, identity, and contamination spectrum of the drug product. Typical quality questions refer to the contamination of parenteral drugs with bacteria or pyrogens derived from bacteria. They may also examine the consistency of individual batches of a vaccine produced from live viruses by inactivation. Measuring contaminations in the active drug ingredient or the final drug product (a pill or a spray) is also an important quality issue.

Given that a drug is efficacious, it can be toxic or non-toxic. Given a drug is non-toxic, it may be efficacious or non-efficacious. It is essential to know that no legislation directly requires animal experiments to demonstrate drug efficacy. Thus, it is possible to bring a drug to the market without any animal experiments for efficacy. For some medicines, animal efficacy studies are not possible, e.g., for human-derived or human-specific proteins that do not have a cognate receptor or target in animals. Animal studies have in the past helped to find drug targets. They were also useful to discover drug candidates, compare them, and gain confidence for their translation to human use. However, testing on humans, i.e., clinical trials, has always been required. Drug candidates that have cost millions to be developed may still fail in clinical trials (for efficacy or safety reasons) despite very promising preclinical data in animals. Therefore, bringing a drug candidate to the market without human trial data is impossible, no matter how many animal experiments are done. Strictly speaking, the animal experiments are not done to achieve drug registration but to prepare the clinical trial. There are other ways to obtain the necessary information that justifies starting a clinical trial, and there are many cases in which these have been used. In the future, the use of animal studies to pave the way for drug trials is likely to decrease.

Safety evaluation of drug candidates follows entirely different rules; usually, there is a defined minimum set of mandatory tests to be performed on animals. These tests are done to ensure the safety of volunteers during clinical trials and reduce the likelihood of long-term adverse effects occurring when large populations are treated. Animal studies can be instrumental in filtering out drugs that may cause acute and severe damage. They also can provide alerts for specific chronic toxicities, developmental and reproductive toxicity, or carcinogenesis. However, many side effects and adversities are only discovered in clinical trials or sometimes in post-marketing surveillance. One reason is that the predictivity of animal models for human adversity is not perfect owing to species-specific differences. Another primary reason is statistical: Animal studies are usually not "powered" to detect rare side effects; it is practically impossible to make animal models sufficiently sensitive for this purpose. For instance, a drug that causes sudden cardiac death in 0.1% of treated humans will kill 1000 patients if given to one million. In a large animal experiment, one may use 100 rats per group (usually less). To observe this rare side effect once, this experiment would need to be performed ten times (which is legally not allowed). And even if one death was found in ten experiments altogether, this finding would be difficult to interpret. A "trick" to overcome this statistical problem is to give animals very high drug doses that would never be given to humans. The assumption is that overdosing would allow the detection of statistically significant effects in smaller groups of animals. This assumption is flawed because chemicals show different activities at very high doses compared to the effective doses. For example, this assumption caused the false (!) rumor that sweeteners like saccharin may be carcinogenic (Ellwein et al., 1990; Lea et al., 2021).

What is the difference between safety and good quality? Safety evaluations are usually only done once in a drug's life cycle. They are meant to derive information on adverse effects of a pure chemical. In general, this is considered to be a compound property that need not be re-evaluated for each production lot[2]. The quality evaluation follows an entirely different logic. It is not about the active drug ingredient itself but the process of producing the final drug product. For example, contaminants may enter the system. Thus, every production run needs testing for as long as new batches are produced. The good quality of one batch does not mean that the next one will have the same quality. A vast number of animals is, e.g., used to test individual batches of botulinum neurotoxin A (used as Botox) in many countries and by many producers. However, the properties of this toxin as such are well-known. Very reliable NAM have been established for quality testing, and they make animal testing redundant in several use domains (Ambrin et al., 2022; Hartung, 2015, 2021).

2.3 Animals in toxicology

Toxicology overlaps with drug research (drug safety) but also has many other aspects. Toxicity evaluations are required for industrial chemicals, pesticides, food ingredients, cosmetics,

[2] Strictly speaking, toxicity is not a chemical property like some physicochemical parameters. The information on toxicity is derived from assays that have noise and uncertainties. Moreover, the extrapolation from model systems to the human population can result in knowledge gaps and errors. Thus, safety information might change if new data become available, e.g., through new test methods or evaluation strategies. This is, however, not a routine process like quality testing.

and many other domains. Toxicological evaluations are not restricted to manufactured classical (pure, low molecular weight) chemicals but are also performed for environmental metabolites (methylmercury), polymers, nanoparticles, and natural products (e.g., algal, fungal, or bacterial toxins) entering the food chain. For most areas, there are still legal requirements that can at present only be fulfilled with animal-based safety studies. A notable exception is cosmetics, for which the use of animal testing is banned in Europe and some other countries. Many modern legislations, like REACH (Registration, Evaluation, Authorisation, and Restriction of Chemicals) in Europe or TSCA (Toxic Substances Control Act) in the US, stipulate the use of non-animal methods where possible. As more and more NAM become the basis of OECD (Organisation for Economic Co-operation and Development) test guidelines, their use instead of animals thus becomes mandatory by law.

What is often not clear enough in public debates is that there is no universal "animal model for toxicity". Instead, many forms of toxicity need to be evaluated one by one in different sets of animal experiments. Toxicity can have many causes and consequences so that no single test method can cover "all toxicities". This situation makes it easier to find alternatives, as there is no need to substitute animals in general, but each type of toxicity test can be targeted, and the panel of animal studies can be substituted step-by-step by NAM. In this context, it is important to understand that animals are only an element (the so-called test system) of many test methods. In analogy, an antibody, a cell culture, or temperature sensors are also test systems of a test method (e.g., a COVID test, a cytotoxicity assay, or a fever monitor). Some test methods that use animals as test systems explore whether a chemical damages DNA (and may thus cause mutations or damage to germ cells). To answer this question, other test systems may also be chosen, e.g., human cells. Other test methods specifically interrogate the capacity of a compound to trigger allergy or eye irritation. Fully valid NAM are available in these areas; thus, animal studies' competitive usefulness (= necessity) is very low. Other test methods broadly cover damage to major organs in adult animals or their offspring. Here, the availability of organs-on-a-chip and microphysiological systems (Marx et al., 2016, 2020) offers new alternatives for replacement (reducing the need for animal testing). The usefulness of animals needs to be questioned for each specific purpose and in the context of the respective test method.

Even though the use of animals may show net benefit, e.g., allergy testing, the final verdict on overall usefulness is decided by the question of whether the animal-based test is better than non-animal alternative approaches (Fig. 4). This is not always the case. For some applications, both animal and NAM-based testing have been shown to be robust, and each method clearly had a net benefit (seen on its own). In such cases, a competitive comparison of robustness and net benefit is required (by law). If the result is that animals are not superior to NAM, then the law prevents animal use, at least in Europe (Directive 2010/63/EU). The rationale is that the ethical costs of NAM are judged to be lower; therefore, animal studies are considered "non-necessary". For instance, it has long been decided by all OECD countries that animals can be entirely replaced for studies on acute toxicities concerning eyes or skin.

A significant change in the field occurred in 2007 when a committee named by the National Academy of Sciences of the USA suggested that animals are of limited use for the future investigation of environmental toxicants (NRC, 2007; Leist et al., 2008; Hartung and Leist, 2008). Similar conclusions were derived later for countermeasures to biological weapons (Hartung and Zurlo, 2012). These ideas have found great resonance also in the European Union, where large research programs have been started to find replacement alternatives for animals in different, more complex toxicological domains (Pallocca et al., 2022b; Moné et al., 2020).

Toxicology is not only concerned with the testing of chemicals but also with understanding how toxicity works and why some substances show certain types of toxicity. This mechanistic research is not only an academic discipline but also plays a crucial role in the industry as so-called "investigative toxicology" (Beilmann et al., 2019). Animals have played an important role in the development of toxicological theory (Leist et al., 2017). For instance, animal models were crucial in identifying carcinogenesis mechanisms or defining the role of xenobiotic metabolism in organ damage. For a long time, animals have been the only model system that has allowed studies on complex toxicological endpoints, concepts, and tissue interactions. With the advent of tissues produced from human cells, advanced microphysiological systems, and organs-on-a-chip, and largely improved analytical sensitivity in human microdosing studies, some powerful alternatives have become available for mechanistic and investigative toxicology (Leist et al., 2017; Marx et al., 2016, 2020; Burt et al., 2020, 2022).

2.4 Understanding of disease (pathogenesis)

Not all medical research is aimed at drug discovery. A large proportion of animals is used in basic biomedical research (Daneshian et al., 2015). Basic pathophysiological concepts are explored in this domain, and animals are used to model aspects of human disease. In more practically oriented forms, this can lead to the discovery of new drug targets, disease biomarkers, or predictors of intervention efficiency. There is also a large, less applied arm of this research direction: Questions focus on which mechanisms may lead to disease onset, modulation, or progression. In many cases, the studies on pathophysiology do not refer to any specific disease. Typical questions refer to more general symptoms or phenomena (e.g., inflammatory responses or neoplastic transformation). They may also address biological regulations (e.g., cell migration, cell stress responses, apoptosis or proteostasis) that are thought to be disturbed in disease. In the past, animal models have contributed valuable knowledge and ideas to this field. More recently, human cell-based systems have been used increasingly, and it has become evident that some of the resources and focus directed to animal experiments may need to be shifted to clinical research and direct investigations of human pathology (Leist and Hartung, 2013; Seok et al., 2013; Suntharalingam et al., 2006).

Also, animal-based pharmacological research is often not di-

rectly linked to developing a specific drug. Instead, significant efforts (and animal numbers) are invested into the basic understanding of pharmacological principles (receptor and signaling systems), how the activity of a drug may be verified (biomarkers of effect), or how the body deals with drugs (e.g., metabolism and excretion). A current example of basic research that does not target new drug development is the examination of the effects of established COVID-19 vaccines, where animal models are used, e.g., to understand how persistently specific immunoglobulins are produced and how the composition of different lymphocyte subsets changes in other body locations. Answers to such questions are of high value and relevance. It is less clear how useful various animal models are in answering such questions; and it is even less clear how often data from animals translate to the human population.

2.5 Basic biological research

One particularly complex (and challenging) area concerning usefulness judgments is basic biological research. However, it is important to mention this area here as it is a major field of animal use (Daneshian et al., 2015; Busquet et al., 2020b). 20-40% of all used experimental animals fall into this area (depending on how much general disease and drug research is counted as "basic research" or "applied research"). Typical questions in basic biology are, e.g., how an animal develops from an oocyte to an adult organism or how this organism copes with environmental changes and aging. These questions arise from a general human curiosity about how the world works, and they are justified by legal frameworks and basic constitutional rights in most countries. Evaluating the usefulness of such research is difficult, if not impossible, but undeniably animal studies have provided a large proportion of the currently available biological knowledge. One driving force has been the possibility to delete genes in the genome of entire animals or in some of their tissues. Technologies initially enabled the study of effects of gene deletion in a knock-out animal model or in cells derived from such an animal rather than producing a knock-out directly in cultured cells (human or rodent). New technologies have recently changed this situation (Driehuis and Clevers, 2017; Hendriks et al., 2020).

Instead of a full usefulness evaluation, one may ask some questions of all future projects: (i) Is the main driver of the project genuinely scientific (or are there other reasons)? (ii) Is the curiosity driven by the interest in the respective process in animals or rather by a motivation to understand the researched phenomenon in humans? (iii) Can the data from animals be used to gain knowledge on humans (given the previous question was answered affirmatively)? (iv) Are there other (animal-free) ways to answer the research question? (v) And does the project have a particularly high level of scientific quality and rigor?

3 Outlook and conclusions

The list of potential applications (use domains) of animal research could be extended nearly endlessly, but that would be beyond the scope of this article. Areas of experimental animal use we have not addressed here comprise research on cosmetics and their ingredients, development of pesticides, testing of medical devices (e.g., hip implants or catheters), development and practice of new surgical techniques, evaluation of cell products, safety testing of food ingredients, testing of weapons and ammunition, evaluating countermeasures to biological and chemical weapons, and development of health-promoting novel food components.

Given the diversity of application domains, the usefulness of animals as a model is hard to judge. The only chance to achieve a consensus between evaluators and arrive at reasonable and robust evaluation outcomes is to focus on sharply delineated areas. The focus here has been on outlining such "use domains" and on pointing out that a complete picture of all advantages and disadvantages should be obtained for the use of animal models for different applications.

A future discussion will address potential metrics of a net benefit. Moreover, the benefit of animal models will need to be judged in the light of potential alternatives. As in many other areas of life, it is the comparative performance that counts. In times when not many alternatives were available, the use of animals could be judged as necessary even though the net benefit as such was relatively moderate. In times of excellent alternatives, very good animal models may not be necessary, i.e., they may not have a comparative advantage. In other words, to keep a ship afloat (staying above the surface water), water depth does not matter.

References

Ambrin, G., Cai, S. and Singh, B. R. (2022). Critical analysis in the advancement of cell-based assays for botulinum neurotoxin. *Crit Rev Microbiol 25*, 1-17. doi:10.1080/1040841X.2022.2035315

Bal-Price, A., Hogberg, H. T., Crofton, K. M. et al. (2018). Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX 35*, 306-352. doi:10.14573/altex.1712081

Beilmann, M., Boonen, H., Czich, A. et al. (2019). Optimizing drug discovery by investigative toxicology: Current and future trends. *ALTEX 36*, 289-313. doi:10.14573/altex.1808181

Burt, T., Young, G., Lee, W. et al. (2020). Phase 0/microdosing approaches: Time for mainstream application in drug development? *Nat Rev Drug Discov 19*, 801-818. doi:10.1038/s41573-020-0080-x

Burt, T., Roffel, A. F., Langer, O. et al. (2022). Strategic, feasibility, economic, and cultural aspects of phase 0 approaches: Is it time to change the drug development process in order to increase productivity? *Clin Transl Sci 15*, 1355-1379. doi:10.1111/cts.13269

Busquet, F., Hartung, T., Pallocca, G. et al. (2020a). Harnessing the power of novel animal-free test methods for the development of COVID-19 drugs and vaccines. *Arch Toxicol 94*, 2263-2272. doi:10.1007/s00204-020-02787-2

Busquet, F., Kleensang, A., Rovida, C. et al. (2020b). New Eu-

ropean 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

Driehuis, E. and Clevers, H. (2017). CRISPR/Cas 9 genome editing and its applications in organoids. *Am J Physiol Gastrointest Liver Physiol 312*, G257-G265. doi:10.1152/ajpgi.00410.2016

Ellwein, L. B. and Cohen, S. M. (1990). The health risks of saccharin revisited. *Crit Rev Toxicol 20*, 311-326. doi:10.3109/10408449009089867

Griesinger, C., Desprez, B., Coecke, S. et al. (2016). Validation of alternative in vitro methods to animal testing: Concepts, challenges, processes and tools. *Adv Exp Med Biol 856*, 65-132. doi:10.1007/978-3-319-33826-2_4

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. and Zurlo, J. (2012). Alternative approaches for medical countermeasures to biological and chemical terrorism and warfare. *ALTEX 29*, 251-260. doi:10.14573/altex.2012.3.251

Hartung, T. (2015). The human whole blood pyrogen test – Lessons learned in twenty years. *ALTEX 32*, 79-100. doi:10.14573/altex.1503241

Hartung, T. (2021). Pyrogen testing revisited on occasion of the 25th anniversary of the whole blood monocyte activation test. *ALTEX 38*, 3-19. doi:10.14573/altex.2101051

Hendriks, D., Clevers, H. and Artegiani, B. (2020). CRISPR-Cas tools and their application in genetic engineering of human stem cells and organoids. *Cell Stem Cell 27*, 705-731. doi:10.1016/j.stem.2020.10.014

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

Lea, I. A., Chappell, G. A. and Wikoff, D. S. (2021). Overall lack of genotoxic activity among five common low- and no-calorie sweeteners: A contemporary review of the collective evidence. *Mutat Res Genet Toxicol Environ Mutagen 868-869*, 503389. doi:10.1016/j.mrgentox.2021.503389

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., Efremova, L. and Karreman, C. (2010). Food for thought ... Considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317. doi:10.14573/altex.2010.4.309

Leist, M. and Hartung, T. (2013). Inflammatory findings on species extrapolations: humans are definitely no 70-kg mice. *Arch Toxicol 87*, 563-567. doi:10.1007/s00204-013-1038-0

Leist, M., Ghallab, A., Graepel, R. et al. (2017). Adverse outcome pathways: Opportunities, limitations and open questions. *Arch Toxicol 91*, 3477-3505. doi:10.1007/s00204-017-2045-3

Marx, U., Andersson, T. B., Bahinski, A. et al. (2016). Biology-inspired microphysiological system approaches to solve the prediction dilemma of substance testing. *ALTEX 33*, 272-321. doi:10.14573/altex.1603161

Marx, U., Akabane, T., Andersson, T. B. et al. (2020). Biology-inspired microphysiological systems to advance patient benefit and animal welfare in drug development. *ALTEX 37*, 365-394. doi:10.14573/altex.2001241

Moné, M. J., Pallocca, G., Escher, S. E. et al. (2020). Setting the stage for next-generation risk assessment with non-animal approaches: The EU-ToxRisk project experience. *Arch Toxicol 94*, 3581-3592. doi:10.1007/s00204-020-02866-4

NRC – National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC, USA: The National Academies Press. http://www.nap.edu/catalog.php?record_id=11970

Pallocca, G., Rovida, C. and Leist, M. (2022a). On the usefulness of animals as a model system (part I): Overview of criteria and focus on robustness. *ALTEX 39*, 347-353. doi:10.14573/altex.2203291

Pallocca, G., Moné, M. J., Kamp, H. et al. (2022b). Next-generation risk assessment of chemicals – Rolling out a human-centric testing strategy to drive 3R implementation: The RISK-HUNT3R project perspective. *ALTEX*, online ahead of print. doi:10.14573/altex.2204051

Schmidt, B. Z., Lehmann, M., Gutbier, S. et al. (2017). In vitro acute and developmental neurotoxicity screening: An overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol 91*, 1-33. doi:10.1007/s00204-016-1805-9

Seok, J., Warren, H. S., Cuenca, A. G. et al. (2013). Inflammation and host response to injury, large scale collaborative research program. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A 110*, 3507-3512. doi:10.1073/pnas.1222878110

Suntharalingam, G., Perry, M. R., Ward, S. et al. (2006). Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *N Engl J Med 355*, 1018-1028. doi:10.1056/NEJMoa063842

Worth, A. P. and Balls, M. (2001). The importance of the prediction model in the validation of alternative tests. *Altern Lab Anim 29*, 135-44. doi:10.1177/026119290102900210

Acknowledgements

This work was supported by BMBF and DFG grants. Support by CEFIC, the Land-BW (NAM-ACCEPT) and funding by the European Union's Horizon 2020 research and innovation programme under grant agreements No 964537 (RISK-HUNT3R), No. 964518 (ToxFree) and No. 825759 (ENDpoiNTs) are acknowledged.