"Everyone is entitled to his own opinion, but not his own facts."
Daniel Patrick "Pat" Moynihan (1927-2003)

Food for Thought ...

Opinion Versus Evidence for the Need to Move Away from Animal Testing

Thomas Hartung

Johns Hopkins University, Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, USA, and University of Konstanz, CAAT-Europe, Konstanz, Germany

Summary

Science is based on facts and their discourse. Willingly or unwillingly, facts are mixed with opinion, i.e., views or judgments formed, not necessarily based on fact or knowledge. This is often necessary, where we have controversial facts or no definitive evidence yet, because we need to take decisions or have to prioritize. Evidence-based approaches aim at identifying the facts and their quality objectively and transparently; they are now increasingly embraced in toxicology, especially by employing systematic reviews, meta-analyses, quality scoring, risk-of-bias tools, etc. These are core to Evidence-based Toxicology. Such approaches aim at minimizing opinion, the "eminence-based" part of science.

Animal experiments are the basis of a lot of our textbook knowledge in the life sciences, have helped to develop desperately needed therapies, and have made this world a safer place. However, they represent only one of the many possible approaches to accomplish all these things. Like all approaches, they come with shortcomings, and their true contribution is often overrated. This article aims to summarize their limitations and challenges beside the ethical and economical concerns (i.e., costs and duration as well as costs following wrong decisions in product development): they include reproducibility, inadequate reporting, statistical under-powering, lack of inter-species predictivity, lack of reflection of human diversity and of real-life exposure. Each and every one of these increasingly discussed aspects of animal experiments can be amended, but this would require enormous additional resources. Together, they prompt a need to engineer a new paradigm to ensure the safety of patients and consumers, new products and therapies.

Keywords: preclinical research, animal models, alternative methods, reproducibility, limitations

1 Introduction

For the 10th anniversary of Food for Thought … in ALTEX, it seemed appropriate to summarize what we have learned on this journey with respect to the core subject of this journal: the need for alternatives to animal experimentation. The series has mostly focused on toxicology, but here the aspects that apply also to drug development and basic research shall be considered.

Sure, we need animal models – when we want to study animals. For example, we have to test drugs for animals in animals. However, when studying human physiology, pharmacology and toxicology, animal models are as much misleading as they are helpful. Half of the results are wrong (Ioannidis, 2005) – we only don't know which half… But the statement "half are wrong" is probably rather optimistic.

2 Evidence versus opinion in toxicology

I am a strong advocate of evidence-based approaches, not least because I was one of the initiators of Evidence-based Toxicology (EBT) and the respective Collaboration and hold the first

Received March 29, 2017
https://doi.org/10.14573/altex.1703291

Chair for Evidence-based Toxicology world-wide, endowed by the Doerenkamp-Zbinden Foundation. These activities aim to bring Evidence-based Medicine to toxicology, i.e., the systematic, objective, and transparent test method assessment and decision-making based on test results. This shall limit bias, prejudice and identify the limits of our knowledge – it is thus exactly the opposite of the shortcuts opinion enables.

Opinion is defined by the Oxford dictionary as "*A view or judgement formed about something, not necessarily based on fact or knowledge*". Hippocrates (ca. 460 – 375 BCE) is quoted "*There are in fact two things, science and opinion; the former begets knowledge, the latter ignorance*." But can we actually avoid opinion in science? The Roman emperor Marcus Aurelius Antonius (121 - 180 CE) stated "*Everything we hear is an opinion, not a fact. Everything we see is a perspective, not the truth*."

I strongly believe that opinion cannot only not be avoided, but is in fact highly valuable as long as we clearly distinguish it from factual evidence and make clear where evidence ends and where opinion starts. First, opinion helps to fill the gaps, for which we have no evidence yet. Expert advice is better than nothing, much better in fact. Second, it is much more entertaining and inspiring: You cannot argue facts, but a good hypothesis – true or not – can spark ideas, controversy, etc. Already as a student, I largely passed on all lectures that were only conveying textbook knowledge (I could learn this for the exams from the textbooks without my faulty notes), but savored those which were spiced with opinion. So, this has been my goal in my talks, lectures and some of my articles.

3 Animal tests are costly and resource intensive

It is difficult to apply economic considerations to all animal experiments in basic research and drug development, as we did for safety testing (Bottini and Hartung, 2009, 2010; Bottini et al., 2007): approaches are so diverse, especially between drug industry and academia, that costs and benefits cannot be contrasted easily.

Toxicological studies become resource intensive for three reasons: (1) they are typically done under Good Laboratory Practice (GLP) quality standards, (2) they treat animals for long periods of time and (3) they assess many endpoints to gain maximum information and avoid missing any harmful effect.

All of this is avoided in other types of research to conserve financial resources, but also because it demands large quantities of material (many kg of test substance) and leads to multiple-testing problems, which decrease the statistical power. In a very simplified view, the economic efficiency of animal tests is determined by whether new, important research is produced or a new drug comes to the market. For both, the impact of a single experiment cannot be judged. This often has more to do with perception than with objective impact. Some people in academia appear to believe that no new line in a textbook can be produced without a new knock-out mouse. We will never know how many wrong decisions are taken in drug development because of misleading animal tests. The performance figures of the few tests analyzed and the drain of the drug pipeline would suggest a substantial number of such mistakes.

Evidence vs. opinion as to economic considerations
While costs and duration of toxicological studies are clearly prohibitive to satisfy societal safety needs, e.g., the often-quoted example of cancer bioassays at $1 million and four years per chemical, this argument is difficult to make for research outside of toxicology.

4 Ethics – where there is an alternative, we must use it instead of harming animals!

Ethical aspects can be left aside here – they should be a no-brainer: It is not only criminal, but no sane person will make animals suffer if there is no need to do so, i.e., an alternative is practically available.

Evidence vs. opinion as to ethical considerations
No evidence is needed if alternatives are available. But whether they are available depends, outside of toxicology where we have formal validation and regulatory acceptance, largely on opinion. It is not realistic to formally validate alternatives for the majority of models in basic research and drug discovery – there are too many models and model variants, and the methods used change too quickly. It is thus critical to shape opinion by informing, teaching the objective assessment of their value, and creating doubt in current practices.

5 Animal experiments are not sufficiently reproducible

They are at least not reproducible enough to work with the group sizes that are typically used. Noteworthy, what is meant by "reproducibility" needs some sharpening here, as it means different things in different disciplines and areas (Goodman et al., 2016). I will choose examples from toxicology, my own "turf", but disease models have been systematically reviewed and summarized before (Hartung, 2013), finding no striking differences.

Arguably, toxicology is an area where we can expect the best reproducibility: Protocols have been standardized over decades into international guidance, much work is done under GLP quality assurance, we use high ("maximum tolerated") doses of substances, and, unlike in pharmacology, we do not have to induce artificial diseases in toxicology. We also pay incredible fees to have the experiments performed by trained professionals: A cancer study in one species for one substance costs $1 million (Basketter et al., 2012), an inhalation study $2.5 million (Hartung, 2016); a developmental neurotoxicity study costs $1.4 million (Smirnova et al., 2014). These are budgets one can only dream of in academia, where our young-

est students do most of the work while "learning on the job", though this does not prohibit us from publishing their work (Hartung, 2013).

The cancer bioassay is a good test case for a reproducibility assessment of animal tests: More than 3,500 studies have been amassed – at today's cost that is $3.5 billion spent. 13% of studies give equivocal results (Seidle, 2006) and the reproducibility was 57% for 121 substances tested repeatedly (Gottman et al., 2001). The OECD guidelines do not make randomization and blinding mandatory, and the guideline statistics do not control for multiple testing, despite the fact that about 60 endpoints are assessed. The cancer bioassay might be a difficult case as some colleagues argue, but when looking at the non-cancer endpoints for 37 substances, very little of the earlier chronic studies was reproduced and consistency between genders and rodent species was low (Wang and Gray, 2015).

What about simpler and shorter animal studies? Severe eye irritation is 70% reproducible (Luechtefeld et al., 2016a). Even validated animal tests do not perform much better: The local lymph node assay for skin allergy is 89% reproducible (Luechtefeld et al., 2016b), and the uterotrophic test for estrogenic endocrine disruption has 26% controversial data if repeated (Browne et al., 2015).

These are only assessments of the reproducibility under the optimal conditions of regulatory guideline studies – this does not say that the results are meaningful for humans. Hallmark papers with respect to non-reproducibility of academic research (Begley and Ellis, 2012; Prinz et al., 2011) have alarmed the scientific community (Macleod, 2011; McGonigle and Ruggeri, 2014; Jarvis and Williams, 2016).

Evidence vs. opinion as to reproducibility

There is increasing evidence that we have a reproducibility problem to which animal experimentation is contributing. Still, more systematic analyses are needed to form a point of reference.

6 Animal experiments are not reported well enough

Efforts to develop guidance on how to report animal studies led to the ARRIVE guidelines (Kilkenny et al., 2010). So, we know what should be reported when writing a scientific paper. When applying this standard and comparing with the reality of 271 randomly picked studies (Kilkenny et al., 2009), the results are more than disappointing: "*Only 59% of the studies stated the hypothesis or objective of the study and the number and characteristics of the animals used. ... Most of the papers surveyed did not use randomisation (87%) or blinding (86%), to reduce bias in animal selection and outcome assessment. Only 70% of the publications that used statistical methods described their methods and presented the results with a measure of error or variability*." More than 300 journals have adopted the ARRIVE guidance, but this seems to be mainly lip-service as first checks suggest (Baker et al., 2014), showing no real improvement in reporting. Notably, these findings apply to the scientific literature, not to the guideline studies used to estimate reproducibility in a type of "best-case scenario" above. A big problem is the generally poor statistics used in publications (Altman, 1998), especially when addressing many effects in the animal at the same time (a chronic toxicity test has 40, a cancer bioassay 60, and a reproductive toxicity study 80 endpoints, without any corrections for multiple testing if using statistics at all). A prominent call for improving reporting of clinical results was published by Landis et al. (2012).

Evidence vs. opinion as to reporting quality

There is clear evidence that reporting standards for animal experiments are not adequate. This does not mean that they are any better for *in vitro* work….

7 Study Design – animal experiments are statistically underpowered, which is compensated by so much standardization that they no longer reflect even their own species

Standardization of animals reduces natural variability and, thus, dramatically reduces the probability of significant findings. We often use inbred strains (genetically "identical twins"), almost always of the same age and gender; in best cases, we randomize for weight differences, etc. We also keep the animals free of any diseases (as "specified pathogen-free") and standardize cages, temperature, and feed. All of this is helpful to improve reproducibility, but our results will also only reflect this exact condition.

The problem that standardization instead impairs reproducibility has been recently discussed (Voelkl and Würbel, 2016). There is ample literature on how these factors impact on results, e.g., strain (Anon, 2009), genetic drift (Papaioannou and Festing, 1980), gender (Clayton and Collins, 2014), cages (Castelhano-Carlos and Baumans, 2009), lack of enrichment of the environment (Wolfer et al., 2004; Würbel, 2007), feed, temperature, diurnal rhythm, time of the year (Kiank et al., 2007), etc. Nevalainen (2014) summarized some of the influential factors, including also seasonal cycle, reproductive cycle, weekend-working day cycle, cage change and room sanitation cycle, diurnal cycle, in-house transport, caging, temperature, humidity, illumination, acoustic environment, odors, cage material, bedding, complexity items, feeding, kinship and humans. They concluded, "*Laboratory animal husbandry issues are an integral but underappreciated part of experimental design, which if ignored can cause major interference with the results*".

In no case was this comprehensively assessed for any given animal test. The reported impacts are anecdotal, it is difficult to say how they jointly impact and how our often-arbitrary choices or lack of control of a parameter distort results. However, it is clear that hardly any experiment shows a general result for a given species.

Evidence vs. opinion as to study designs
There is clear qualitative evidence that many impacting factors are either not controlled or standardized to an extent that results are no longer generally applicable. There is no quantitative evidence for most of these factors though. Opportunities to remedy these problems are limited by feasible group sizes, as most designs are already underpowered.

8 Animal experiments do not even predict other animal species

I am often quoted for the rather simple statement "Humans are not 70 kg rats!" (Hartung, 2009). But rats are also not 300 g mice! The difference here is that we can compare because some highly standardized (toxicological) tests are being done on more than one species (Leist and Hartung, 2013). The results are discouraging: mice and rats predict each other for carcinogenicity of chemicals by 57% (Gray et al., 1995), and this value drops if we also look for prediction of the target organ that is affected (Gold et al., 1991). Rats and rabbits (as well as other species) predict reproductive toxicity of each other by 60% (Bailey et al., 2005). Guinea pigs and mice predict skin sensitization of each other in 77% of cases (Luechtefeld et al., 2016b). Mouse and rat have little prediction for each other's chronic toxicities (Wang and Gray, 2015).

There is no reason to assume that any species predicts effects in humans any better (Perlman, 2016) than it predicts effects in another animal species. Hardly any species comparisons have been done for basic research and drug discovery. However, often even differences between mouse strains are reported.

Evidence vs. opinion as to inter-species predictivity
There is clear evidence for tremendous species differences from toxicology, but this is limited for other areas of research. There is no reason to assume that toxicology has more inter-species variances; on the contrary, here substance effects are studied at high doses, most substances act in a manner that is not receptor-mediated, and, unlike in pharmacology, there is no additional complication of a disease model, in which the substance is tested for modulatory effects.

9 Animal experiments do not reflect human diversity, exposure, and treatment

The lack of natural diversity in our animal experiments was already addressed. Humans are different from inbred mice in many aspects: our weights, our age, our lifestyle, our genetics, our history of diseases cover broad ranges. This all makes it very difficult to predict substance effects, even more if one is trying to treat diseases that are at different stages in combination with different comorbidities and other parallel treatments. This has nothing to do with the monotreatments in standardized disease models. For a list of differences, see Hartung (2013).

With respect to toxic effects, we typically study the acute and local effects of high doses in animal experiments, relevant if at all in workplace situations. However, for general human health, we should be concerned about low and chronic exposures. We are exposed to mixtures of chemicals in and from the different products. Differences in the kinetics and metabolism of substances add to the problem. The human organism often varies dramatically from the animal with respect to uptake, distribution and excretion of substances, and forms very different metabolites of the same substance.

Evidence vs. opinion as to not reflected human diversity
There is no doubt about this, but also not any answer showing how to tackle the problem. Panels of human cells representing diverse individuals would work only in a few cases.

10 What can we improve in our animal experiments?

Table 1 shows a personal scoring for the available evidence:

Tab. 1: Strength of evidence for limitations of animal tests

Limitations	Toxicology (guideline studies)	Drug Discovery and Basic Research
Economic – financial	+++	+ – ++
Economic – duration	+++	+ – ++
Economic – test substance need	+ – +++	+ – ++
Ethical (depending on information need and availability of alternatives)	0 – +++	0 – +++
Reproducibility	+++	+++ (assumed to be similar to tox.)
Reporting quality	0 (GLP) – +++	+++
Study design (uncontrolled and arbitrarily standardized parameters)	++ (multiple testing)	+ – ++
Inter-species predictivity	++	++
Human diversity not reflected	+++	+++

Sure, we can improve many aspects of how we do our animal tests (leaving aside all the aspects of reducing distress and suffering of the animals (Zurlo and Hutchinson, 2014)): We can use more genetically diverse animals in enriched environments, study both genders and several species. Richter et al. (2011)

have actually shown that systematic variation of experimental parameters improves reproducibility.

We can analyze the kinetics of substances in different species, including man, and improve our extrapolation to humans (Bale et al., 2014; Tsaioun et al., 2016), especially by integrating information from *in vitro* epithelial barrier models (Gordon et al., 2015). We can use and properly report the right statistics, which in turn will strongly increase the necessary animal group sizes. However, all this would make our experiments incredibly expensive.

We can argue that from a given budget it is better to publish fewer but more meaningful results. We can also standardize and validate further animal tests – this would improve the comparability of results and show more clearly the strengths and weaknesses of these models, but again these are lengthy and costly exercises that would likely produce many disappointments about broadly used models.

The traditional way of handling this in toxicology is the safety or assessment factor, i.e., the (no) effect level of a substance is corrected by a factor, typically 10, for possible inter-individual and another factor, typically 10, for inter-species differences. Often additional factors are added if further limitations exist in the data. It is pragmatic to err on the side of safety.

The problem is that such uncertainty factors cannot be modelled *in vitro* or *in silico*. They also come on top of additional safety measures, such as choosing the most sensitive species and using high (maximum tolerated) doses. However, neither assessment factors nor high doses help in disease and drug effect models.

11 What can we do instead of animal experiments?

First, we should study what we can in humans in order to understand human physiology, disease and treatment. We do not really take enough advantage of the ongoing daily exposure of people, i.e., epidemiology, though advances like biomonitoring, biomarkers, biobanking, and the human exposome (Escher et al., 2017) must be noted. Also, microdosing of substances in humans and more comprehensive assessments when first going into humans represent some, though limited, opportunities (Seymour, 2009).

Human tissues and their reconstruction *in vitro* (Alépée et al., 2014), including bioprinting and organ-on-chip bioengineering, represent the next line of opportunities (Andersen et al., 2014; Marx et al., 2016). The current paradigm shift toward organotypic cultures with organ architectures and organ functionalities presents avenues to more meaningful models.

These prospects should not blind us to the shortcomings of these models and the challenges ahead (Hartung, 2007, 2013; Pamies et al., 2017; Pamies and Hartung, 2017). Their quality assurance, e.g., Good Cell Culture Practice and *in vitro* reporting standards (Coecke et al., 2005; Leist et al., 2010; Pamies et al., 2017), are only under development. Incredibly high rates of misidentified cells, mycoplasma infections, and genetic aberration in culture challenge this part of research no less than the animal tests discussed here. They will be part of a roadmap to a more comprehensive coverage of human hazards by new approach methods (Basketter et al., 2012; Leist et al., 2014).

12 Conclusions

This discussion of the shortcomings of first of all animal tests is not a call to abandon them right now. Information that is often not right might still be better than no information at all. It means, however, that in light of these limitations we need really good justification to harm an animal. Only looking for and openly discussing the limitations of an individual animal test will enable us to move forward, and often this means away from the animal model. In some cases, we might not yet have an alternative, but it is important to identify the goal of creating one.

The critical step is to understand the strengths and limitations of our models, both *in vivo* and *in vitro*. The systematic assessment of study quality (Samuel et al., 2016) is a key step toward such analysis, favorably by systematic reviews (Stephens et al., 2016), as recently proposed for example for endocrine disrupting chemicals (Vandenberg et al., 2016). Then we can start combining them to move towards more meaningful results in integrated testing strategies (Hartung et al., 2013; Rovida et al., 2015).

This all is more easily said than done. Science has too few self-critical and self-controlling mechanisms. Nobody writes more than is absolutely necessary about the weaknesses of the models in scientific papers or grant applications. Those who are more careful and control their models and results are penalized, as they cannot publish as quickly and as much exciting stuff as their colleagues.

But these exciting results are often shaky – the only 10-25% reproducibility of important scientific papers published by the pharmaceutical industry is alarming (Begley and Ellis, 2012; Prinz et al., 2011). Science works by forgetting the irreproducible results over time – we stop citing them. However, the growth of the scientific community and the ever-easier access to literature allows these studies to resurface again and again, cited by those who don't know any better. Given the lottery of our peer-review system and the overload of the experts with review duties, and the business models of publishers, which in the end make everything publishable somewhere, we should not be surprised about the increasingly perceived "reproducibility crisis" (Baker, 2016).

The efforts by NIH and others to address this are laudable, but in the end, we need a "scientific enlightenment movement", a type of restart as Life Science 2.0. Clinical research has started this with Evidence-based Medicine. We need something similar in preclinical research. Efforts of systematic reviews of animal studies (Ritskes-Hoitinga et al., 2014) or Evidence-based Toxi-

cology (Hoffmann and Hartung, 2006; Hoffmann et al., 2014) represent such starting points. Together with the emerging technologies for a toxicology for the 21st century, they represent a framework for systematically developing safety sciences (Busquet and Hartung, 2017).

References

Alépée, N., Bahinski, T., Daneshian, M. et al. (2014). State-of-the-art of 3D cultures (organs-on-a-chip) in safety testing and pathophysiology – a t4 report. *ALTEX 31*, 441-477. https://doi.org/10.14573/altex1406111

Altman, D. G. (1998). Statistical reviewing for medical journals. *Stat Med 17*, 2661-2674. https://doi.org/10.1002/(SICI)1097-0258(19981215)17:23<2661::AID-SIM33>3.0.CO;2-B

Andersen, M., Betts, K., Dragan, Y. et al. (2014). Developing microphysiological systems for use as regulatory tools – challenges and opportunities – extended online version. *ALTEX 31*, 364-367. https://doi.org/10.14573/altex.1405151

Anon (2009). Troublesome variability in mouse studies. *Nat Neurosci 12*, 1075-1075. https://doi.org/10.1038/nn0909-1075

Bailey, J., Knight, A. and Balcombe, J. (2005). The future of teratology research is in vitro. *Biogenic Amines 19*, 97-145. https://doi.org/10.1163/1569391053722755

Baker, D., Lidster, K., Sottomayor, A. and Amor, S. (2014). Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol 12*, e1001756-e1001756. https://doi.org/10.1371/journal.pbio.1001756

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature 533*, 452-454. https://doi.org/10.1038/533452a

Bale, A. S., Kenyon, E., Flynn, T. J. et al. (2014). Correlating in vitro data to in vivo findings for risk assessment. *ALTEX 31*, 79-90. https://doi.org/10.14573/altex.1310011

Basketter, D. A., Clewell, H., Kimber, I. et al. (2012). A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. *ALTEX 29*, 3-89. https://doi.org/10.14573/altex.2012.1.003

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature 483*, 531-533. https://doi.org/10.1038/483531a

Bottini, A. A., Amcoff, P. and Hartung, T. (2007). Food for thought … on globalization of alternative methods. *ALTEX 24*, 255-261. https://doi.org/10.14573/altex.2007.4.255

Bottini, A. A. and Hartung, T. (2009). Food for thought … on economics of animal testing. *ALTEX 26*, 3-16. https://doi.org/10.14573/altex.2009.1.3

Bottini, A. A. and Hartung, T. (2010). The economics of animal testing. *ALTEX 27, Spec Issue*, 67-77. http://www.altex.ch/resources/EMS.pdf

Browne, P., Judson, R. S., Casey, W. M. et al. (2015). Screening chemicals for estrogen receptor bioactivity using a computational model. *Environ Sci Technol 49*, 8804-8814. https://doi.org/10.1021/acs.est.5b02641

Busquet, F. and Hartung, T. (2017). The need for strategic development of safety sciences. *ALTEX 34*, 3-21. https://doi.org/10.14573/altex.1701031

Castelhano-Carlos, M. J. and Baumans, V. (2009). The impact of light, noise, cage cleaning and in-house transport on welfare and stress of laboratory rats. *Lab Anim 43*, 311-327. https://doi.org/10.1258/la.2009.0080098

Clayton, J. A. and Collins, F. S. (2014). NIH to balance sex in cell and animal studies. *Nature 509*, 282-283. https://doi.org/10.1038/509282a

Coecke, S., Balls, M., Bowe, G. et al. (2005). Guidance on good cell culture practice. *Altern Lab Anim 33*, 261-287.

Escher, B. I., Hackermüller, J., Polte, T. et al. (2017). From the exposome to mechanistic understanding of chemical-induced adverse effects. *Environ Int 99*, 97-106. https://doi.org/10.1016/j.envint.2016.11.029

Gold, L. S., Slone, T. H., Manley, N. B. and Bernstein, L. (1991). Target organs in chronic bioassays of 533 chemical carcinogens. *Environ Health Perspect 93*, 233-246. https://doi.org/10.1289/ehp.9193233

Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci Transl Med 8*, 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Gordon, S., Daneshian, M., Bouwstra, J. et al. (2015). Non-animal models of epithelial barriers (skin, intestine and lung) in research, industrial applications and regulatory toxicology. *ALTEX 32*, 327-378. https://doi.org/10.14573/altex.1510051

Gottmann, E., Kramer, S., Pfahringer, B. and Helma, C. (2001). Data quality in predictive toxicology: Reproducibility of rodent carcinogenicity experiments. *Environ Health Perspect 109*, 509-514. https://doi.org/10.1289/ehp.01109509

Gray, G. M., Li, P., Shlyakhter, I. and Wilson, R. (1995). An empirical examination of factors influencing prediction of carcinogenic hazard across species. *Regul Toxicol Pharmacol 22*, 283-291. https://doi.org/10.1006/rtph.1995.0011

Hartung, T. (2007). Food for thought … on cell culture. *ALTEX 24*, 143-147. https://doi.org/10.14573/altex.2007.3.143

Hartung, T. (2009). Toxicology for the twenty-first century. *Nature 460*, 208-212. https://doi.org/10.1038/460208a

Hartung, T. (2013). Look back in anger – what clinical studies tell us about preclinical work. *ALTEX 30*, 275-291. https://doi.org/10.14573/altex.2013.3.275

Hartung, T., Luechtefeld, T., Maertens, A. and Kleensang, A. (2013). Integrated testing strategies for safety assessments. *ALTEX 30*, 3-18. https://doi.org/10.14573/altex.2013.1.003

Hartung, T. (2016). E-Cigarettes and the need and opportunities for alternatives to animal testing. *ALTEX 33*, 211-224. https://doi.org/10.14573/altex.1606291

Hartung, T. (2017b). Utility of the adverse outcome pathway concept in drug development. *Expert Opin In Drug Metab Toxicol 13*, 1-3. https://doi.org/10.1080/17425255.2017.1246535

Hoffmann, S. and Hartung, T. (2006). Towards an evidence-

based toxicology. *Human Exp Toxicol 25*, 497-513. https://doi.org/10.1191/0960327106het648oa

Hoffmann, S., Stephens, M. and Hartung, T. (2014). Evidence-based toxicology. In P. Wexler (ed.), *Encyclopedia of Toxicology*. 3rd edition, Vol. 2 (565-567). Elsevier Inc., Academic Press. https://doi.org/10.1016/B978-0-12-386454-3.01060-5

Ioannidis, J. P. A. J. (2005). Why most published research findings are false. *PLoS Med 2*, e124-e124. https://doi.org/10.1371/journal.pmed.0020124

Jarvis, M. F. and Williams, M. (2016). Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends Pharmacol Sci 37*, 290-302. https://doi.org/10.1016/j.tips.2015.12.001

Kiank, C., Koerner, P., Kessler, W. et al. (2007). Seasonal variations in inflammatory responses to sepsis and stress in mice. *Crit Care Med 35*, 2352-2358. https://doi.org/10.1097/01.CCM.0000282078.80187.7F

Kilkenny, C., Parsons, N., Kadyszewski, E. et al. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One 4*, e7824-e7824. https://doi.org/10.1371/journal.pone.0007824

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

Landis, S. C., Amara, S. G., Asadullah, K. et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature 490*, 187-191. https://doi.org/10.1038/nature11556

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. https://doi.org/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. https://doi.org/10.1007/s00204-013-1038-0

Leist, M., Hasiwa, N., Rovida, C. et al. (2014). Consensus report on the future of animal-free systemic toxicity testing. *ALTEX 31*, 341-356. https://doi.org/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. https://doi.org/10.14573/altex.1510053

Luechtefeld, T., Maertens, A., Russo, D. P. et al. (2016b). Analysis of publically available skin sensitization data from REACH registrations 2008-2014. *ALTEX 33*, 135-148. https://doi.org/10.14573/altex.1510055

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

Macleod, M. (2011). Why animal research needs to improve. *Nature 477*, 511-511. https://doi.org/10.1038/477511a

McGonigle, P. and Ruggeri, B. (2014). Animal models of human disease: Challenges in enabling translation. *Biochem Pharmacol 87*, 162-171. https://doi.org/10.1016/j.bcp.2013.08.006

Nevalainen, T. (2014). Animal husbandry and experimental design. *ILAR J 55*, 392-398. https://doi.org/10.1093/ilar/ilu035

Pamies, D., Bal-Price, A., Simeonov, A. et al. (2017). Good cell culture practice for stem cells and stem-cell-derived models. *ALTEX 34*, 95-132. https://doi.org/10.14573/altex.1607121

Pamies, D. and Hartung, T. (2017). 21st century cell culture for 21st century toxicology. *Chem Res Toxicol 30*, 43-52. https://doi.org/10.1021/acs.chemrestox.6b00269

Papaioannou, V. E. and Festing, M. F. (1980). Genetic drift in a stock of laboratory mice. *Lab Anim 14*, 11-13. https://doi.org/10.1258/002367780780943015

Perlman, R. L. (2016). Mouse models of human disease: An evolutionary perspective. *Evol Med Public Health 1*, 170-176. https://doi.org/10.1093/emph/eow014

Prinz, F., Schlange, T. and Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nat Rev Drug Discov 10*, 712. https://doi.org/10.1038/nrd3439-c1

Richter, S. H., Garner, J. P., Zipser, B. et al. (2011). Effect of population heterogenization on the reproducibility of mouse behavior: A multi-laboratory study. *PLoS One 6*, e16461. https://doi.org/10.1371/journal.pone.0016461

Ritskes-Hoitinga, M., Leenaars, M., Avey, M. et al. (2014). Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. *Cochrane Database Syst Rev 3*, ED000078. https://doi.org/10.1002/14651858.ED000078

Rovida, C., Alépée, N., Api, A. M. et al. (2015). Integrated testing strategies (ITS) for safety assessment. *ALTEX 32*, 171-181. https://doi.org/10.14573/altex.1506201

Samuel, G. O., Hoffmann, S., Wright, R. et al. (2016). Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. *Environ Int 92-93*, 630-646. https://doi.org/10.1016/j.envint.2016.03.010

Seidle, T. (2006). Chemicals and cancer: What the regulators won't tell you. PETA Europe Ltd. http://www.peta.nl/pdf/PetaCancerReport.pdf

Seymour, M. (2009). The best model for humans is human – how to accelerate early drug development safely. *Altern Lab Anim 37, Suppl 1*, 61-65.

Smirnova, L., Hogberg, H. T., Leist, M. and Hartung, T. (2014). Developmental neurotoxicity – challenges in the 21st century and in vitro opportunities. *ALTEX 31*, 129-156. https://doi.org/10.14573/altex.1403271

Stephens, M. L., Betts, K., Beck, N. B. et al. (2016). The emergence of systematic review in toxicology. *Toxicol Sci 52*, 10-16. https://doi.org/10.1093/toxsci/kfw059

Tsaioun, K., Blaauboer, B. J. and Hartung, T. (2016). Evi-

dence-based absorption, distribution, metabolism, excretion and toxicity (ADMET) and the role of alternative methods. *ALTEX 33*, 343-358. https://doi.org/10.14573/altex.1610101

Vandenberg, L. N., Ågerstrand, M., Beronius, A. et al. (2016). A proposed framework for the systematic review and integrated assessment (SYRINA) of endocrine disrupting chemicals. *Environ Health 15*, 74. https://doi.org/10.1186/s12940-016-0156-6

Voelkl, B. and Würbel, H. (2016). Reproducibility crisis: Are we ignoring reaction norms? *Trends Pharmacol Sci 37*, 509-510. https://doi.org/10.1016/j.tips.2016.05.003

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

Wolfer, D. P., Litvin, O., Morf, S. et al. (2004). Laboratory animal welfare: Cage enrichment and mouse behaviour. *Nature 432*, 821-822. https://doi.org/10.1038/432821a

Würbel, H. (2007). Environmental enrichment does not disrupt standardisation of animal experiments. *ALTEX 24*, *Spec Issue*, 70-73. http://www.altex.ch/All-issues/Issue.50.html?iid=88

Zurlo, J. and Hutchinson, E. (2014). The state of animal welfare in the context of refinement. *ALTEX 31*, 4-10. https://doi.org/10.14573/altex.1312191

Conflict of interest

As the author refers to organotypic cell cultures, he would like to declare his interests as founder of Organome LLC, Baltimore, consultant to AstraZeneca and Chief Scientific Advisor/co-owner of Atheralabs, Luxembourg.

Acknowledgement

The author would like to thank Drs Malcolm MacLeod and Ulrich Dirnagl, as well as two anonymous reviewers, for their comments on an earlier version of this manuscript.

Correspondence to

Thomas Hartung, MD PhD
JHU Bloomberg School of Public Health, CAAT
W7032
615 N. Wolfe St.
Baltimore, MD 21224, USA
Phone: +1 410 614 4990
e-mail: THartun1@jhu.edu