Evolution of toxicological science: the need for change

Thomas Hartung

Center for Alternatives to Animal Testing,
Johns Hopkins Bloomberg School of Public Health,
615 North Wolfe Street,
W7032, Baltimore, MD 21205, USA
and
CAAT-Europe,
University of Konstanz, Germany
Email: thartung@jhsph.edu

Abstract: Evolution requires change but the field of toxicology has not kept pace. The top ten problems urging change are: 1) disparity of testing requirements and risk acceptance for different products and geographical areas; 2) throughput and testing costs versus testing needs; 3) limited predictivity for humans; 4) precautionary approaches from drug development adapted to other areas; 5) animal use; 6) traditional tests unsuitable for new products; 7) lack of coverage for new hazards; 8) failure to address mixtures of toxicants; 9) lack of coverage for individual susceptibilities and vulnerable subpopulations; 10) poor basic research and publication standards. The intransigence at the root of these problems is discussed with reference to current international toxicological policies and methods. For each, the limitations are reviewed with reference to key literature. While current approaches are still needed, there is room for substantial change. To meet the challenges of the 21st century, revolution rather than evolution is required.

Keywords: toxicity testing; high-throughput; precautionary approach; mixtures; susceptible populations.

Reference to this paper should be made as follows: Hartung, T. (2017) 'Evolution of toxicological science: the need for change', *Int. J. Risk Assessment and Management*, Vol. 20, Nos. 1/2/3, pp.21–45.

Biographical notes: Thomas Hartung is a Professor of Toxicology (Chair for Evidence-based Toxicology), Pharmacology, Molecular Microbiology and Immunology at Johns Hopkins Bloomberg School of Public Health, Baltimore, and University of Konstanz, Germany. He also is the Director of their Centers for Alternatives to Animal Testing (CAAT).

This paper is a revised and expanded version of a paper entitled 'Evolution of toxicological science: the need for change' presented at Risk Science International, Ottawa, 4–6 March 2013.

> "I cannot say whether things will get better if we change; what I can say is they must change if they are to get better." (Georg Christoph Lichtenberg, 1742–1799)

Copyright © The Author(s) 2016. Published by Inderscience Publishers Ltd. This is an Open Access Article distributed under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1 Introduction

Toxicology has changed both in real terms with a chemical revolution – the introduction in the 19th and 20th centuries of thousands of synthetic chemicals – and also with increased awareness of multiple and mixed exposures to chemicals arising from food and food packaging as well as numerous other sources. Approaches that were once able to describe toxicity of the relatively high exposures addressed by Paracelsus, or the later occupational exposures described in Hunter's 1955 classic *Diseases of Occupations*, have had to evolve to determine the toxicities of lower exposures. Even so, the current approaches appear inadequate for determining the subtler nature of present day exposures.

The title of this article (invited for the *Risk in the 21st Century* seminar in Ottawa, 2013) is a contradiction in itself – evolution implies continuous change. So if there is need for change now, it is because the evolutionary process did not work adequately. In the words of Charles Darwin, "It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change." If we make a case for change here, it is because toxicology has not always been embracing it.

This author has argued elsewhere that a scientific revolution, more so than an evolution, is required – and likely to take place – in toxicology (Hartung, 2008c). The argument was based Thomas Samuel Kuhn's classic *The Structure of Scientific Revolutions* (see Table 1).

Table 1 Key assumptions of Thomas Kuhn's *The Structure of Scientific Revolutions* summarised by Frank Pajares, Emory University

1	A scientific community cannot practice its trade without some set of received beliefs.
2	The nature of the 'rigorous and rigid' preparation helps ensure that the received beliefs exert a 'deep hold' on the student's mind
3	Normal science "is predicated on the assumption that the scientific community knows what the world is like" – scientists take great pains to defend that assumption; to this end, "normal science often suppresses fundamental novelties because they are necessarily subversive of its basic commitments."
4	Research is "a strenuous and devoted attempt to force nature into the conceptual boxes supplied by professional education."
5	A shift in professional commitments to shared assumptions takes place when an anomaly "subverts the existing tradition of scientific practice"; these shifts are what Kuhn describes as scientific revolutions – "the tradition-shattering complements to the tradition-bound activity of normal science."

Source: Pajares, F. (n.d.)

It seems quite evident that much of this can be easily applied to toxicology. The anomalies currently subverting the field are, among others:

a awareness of how many substances have never been tested and the resulting legislation to address this lack of testing (REACH, the emerging TSCA reauthorisation, etc.) (Hartung, 2010d) b the shift in pharma industry to new entities, especially biologicals (such as human proteins and antibodies) (Rovida et al., 2015) and medical countermeasures for biological and chemical warfare and terrorism (Hartung and Zurlo, 2012), but also nanoparticles (Hartung, 2010c; Hartung and Sabbioni, 2011), which inevitably require methodological changes

c the crisis resulting from the fact that fewer and fewer therapeutic agents make it to the market, with concurrent demands for reviews to determine whether the candidate compounds were correctly selected for development (Hartung, 2013)

d market forces have entered the field of alternatives and the first methods are already resulting in turnovers of tens or even hundreds of millions of euros, with lobbyists creating new pressures (Bottini and Hartung, 2009a)

e legislation forestalling scientific developments – the most pertinent example being the 7th Amendment of the EU Cosmetics Directive, whereby animal tests are to be banned – even in the absence of alternatives – in order to create (via legislation) the necessary pressure on industry to develop the alternative approaches which are currently lacking (Hartung, 2008a)

f globalisation (Bottini et al., 2007): global markets, global companies, the need for the global harmonisation of regulations, new information technologies....

Whether evolutionary or revolutionary, there are many to update the toolbox of toxicology. The top ten problems urging for change are:

1 disparity of testing requirements and risk acceptance for different products and geographical areas

2 throughput and costs of testing versus testing needs

3 limited predictivity for humans

4 precautionary approaches from drug development adapted to other areas

5 animal use

6. new products not suitable for traditional tests

7 new hazards not covered

8 mixtures of toxicants not addressed

9 individual susceptibilities and vulnerable subpopulations not covered

10 poor basic research and publication standards.

Points 1 through 4 are more application-based, whereas points 5 to 9 are methodological shortcomings, and point 10 represents a more general limitation of today's scientific approaches. In the following, these aspects will be elaborated in some detail, primarily within traditional regulatory toxicology which is based predominantly on animal tests. Discussion of these challenges, it should be noted, does not mean that we necessarily have better tools at hand. Many of these challenges hold true for the available alternatives as well, and for other problems there are no alternative approaches available. All methods have their limitations, and we need to remain conscious of these and the continuing need to improve the toolbox and its applications.

2 The needs for change

2.1 The disparity of testing requirements and risk acceptance for different products and geographical areas

The level and extent of testing is very different for various products: while pesticides, for example, are subjected to more than 30 animal tests, food additives often are not tested at all (Neltner et al., 2013). Sure, pesticides are designed to kill (insects, etc.) and food additives are designed to be safe, but the latter holds true also for pharmaceuticals and we see how often it does not work out. At the same time, it is quite difficult to request safety standards for food additives, which many of the 'natural' ingredients do not meet: For example, 23 of 32 tested chemicals in roasted coffee (72%) were positive in the cancer bioassay in rats (Ames and Gold, 2000; Gold et al., 2005). But why are testing demands for animal feed often more extensive than human ones? Why can thresholds of toxicological concern, as a very pragmatic exposure consideration, be used for food contaminants but not for cosmetic ingredients?

The opposite problem, however, can hold true too, i.e., that methods from one sector are simply applied to another when they do not necessarily fit. Key examples are the many methods developed in the context of drug development now applied to chemicals in the context of REACH. In the first case, substances have no commercial value and it is usually possible to change the lead substance with others, while in the latter, substances often used in thousands of products for decades with the respective safety experience are incriminated.

There is also a disparity of testing requirements in different economical regions. We have argued that global industries need global safety standards (Bottini et al., 2007). Science is international and the more science-based our safety assessments are, the more harmonised they will be. Local standards of 'traditional approaches' impede the introduction of new approaches in all countries or regions, as their corporate sector will keep the old standards until the last important region accepts the new. Notably, some international trends are the exact opposite; when, for example, Europe bans testing of cosmetics and their ingredients (Hartung, 2008a) and the US Safe Cosmetics Act may introduce considerable testing demands (Knight and Rovida, 2014). The Safe Cosmetics and Personal Care Products Act of 2013 (H.R. 1385), introduced in the US House of Representatives on March 21, 2013, will ensure the safety of personal care products by assessing cosmetics ingredients and phasing out the most harmful substances (i.e., those suspected to cause cancer or reproductive/other adverse health effects) (Knight and Rovida, 2014). Reisinger et al. (2015) make the case for the example of skin sensitisation.

Taken together, more harmonised testing requirements across industrial sectors – but tailored to the actual use and risk – represent needed changes in toxicity assessments.

2.2 Throughput and costs of testing versus testing needs

Worldwide regulatory decisions for product safety of goods traded at $10 trillion per year are predominantly made on the basis of animal testing (Bottini and Hartung, 2009, 2010).

A reasonable estimate is that people are exposed to about 100,000 relevant synthetic chemicals (84,000 are listed in the cumulative US TSCA inventory, 100,000 in the EU EINECS inventory) in contrast to the 5,000 to 10,000 (Luechtefeld et al., 2016) for which actual (widely varying in depth) safety assessments exist. The knowledge gap is tremendous. We are most likely exposed to an even larger number of chemicals, however, given all naturally occurring sources. A single plant extract used as a drug, for example, can contain 40,000 substances. With regard to possible toxic properties, there is no difference between a substance produced by chemical synthesis or by the metabolism of an organism – on the contrary, some of the most toxic substances are 'natural' because evolution has optimised these poisons. Similarly, there are byproducts from chemical syntheses to be considered. Petrochemicals fall somewhere in-between, as they are very heterogeneous as natural products to start with and fractionation will always remain partial (Patlewicz et al., 2014). Thus, in principle, testing needs are enormous.

At costs of $5 to 20 million for a full assessment of a high-production volume chemical, drug, or pesticide, however, we can hardly apply traditional methods to such enormous numbers of substances. We have earlier shown what it means to request extensive testing for the existing chemicals under REACH (Hartung and Rovida, 2009; Rovida and Hartung, 2009): excessive costs, large numbers of animals, and feasibility problems if the current guidance was applied to the letter. The testing proposals offered by industry in response are very different (Rovida et al., 2011), and the extent of actual testing finally taking place is an open question.

The costs, duration, and throughput of traditional methods are also major roadblocks for early safety assessments in product development. The pharmaceutical industry has spearheaded this under the label of *front-loading of toxicology*. More recently, the chemical industry embraced the concept as *green toxicology* (Maertens et al., 2014). This is the application of predictive toxicology to chemicals with the specific intent of improving their design for hazard reduction. The 12 principles of green chemistry outline a strategy to reduce hazard through molecular and process design. Reducing toxicity is at the core of green chemistry and sustainability, therefore the input of toxicologists early in the chemical enterprise is essential to inform the choices of molecular designers in selecting less hazardous design strategies. Information derived from the combination of mechanistic and computational toxicology forms the nexus between toxicology and green chemistry. Each group is trained to examine, understand, and describe aspects of the structure/hazard relationship from a narrow perspective. We need collaboration among scientists working in complementary fields to discover common ground in the quest for safer chemicals. Such advances require *in vitro* and *in silico* tools that are cost effective and fast. The major opportunity lies in the fact that they do not need ultimate validation for regulatory acceptance. Most recently, there is a lot of advance of read-across approaches (Patlewicz et al., 2014), which are ideally suited for this purpose. The recent move to Good Read-Across Practice guidance (Ball et al., 2016) and availability of large datasets on *in vivo* toxicology (Luechtefeld et al., 2016), will boost this field. Helping to choose substances with a better likelihood of being non-hazardous is sufficiently attractive to foster dialogue between synthetic chemists and toxicologists. This is a communication and workflow challenge (though toxicology is one of the most chemical-structure-oriented of the medical sciences). From the toxicologist's standpoint, it requires offering suitable methods despite their associated uncertainties.

Taken together, only faster and more cost-efficient methods will be able to match the pace of introduction of new substances and allow safety considerations at earlier stages of product development to handle the backlog of old, not-appropriately-tested substances.

2.3 Limited predictivity for humans

Limited predictivity of animal studies is not only toxicology's problem. Recent systematic reviews of the predictive value of animal models have not been favourable (Roberts et al., 2002; Pound et al., 2004; Hackam and Redelmeier, 2006; Perel et al., 2007; Hackam, 2007; van der Worp et al., 2010). There are now even more or less 'systematic' reviews of the systematic reviews (Pound et al., 2004; Mignini and Khan, 2006; Knight, 2007; Briel et al., 2013), showing that there is room for improvement.

Drugs undergo extensive testing in animals both for efficacy and toxicity. But are we sorting out the wrong candidate substances? Substances are, for example, sorted out because of mutagenicity findings in tests in which kitchen salt and sugar would fail (Pottenger et al., 2007). Even aspirin likely would fail the preclinical stage today (Hartung, 2009c). The predictive value of animal testing in general – and for toxicology in particular – can best be estimated from the drug development process: in order to arrive, finally, at a safe and efficient drug – from drug discovery to FDA approval – takes an average of 10 to 15 years and costs more than $1 billion (Mundae and Östor, 2010; Tamimi and Ellis, 2010; Gilbert et al., 2003). In some estimates, when the costs of failed prospective drugs are factored in, the cost of a single drug development has soared from $1.1 billion in 1995 to $1.7 billion in 2002. This has continued with an average in 2012 of $4 billion and up to $11 billion (quoted by *Forbes*) for a successful launch to the market.[1] The figures are not very different for biopharmaceuticals compared to small molecules (DiMasi and Grabowski, 2007). A recent analysis over the last 60 years showed the continuous decline of substances making it to the market per $billion spent (Scannell et al., 2012), even when corrected for inflation. The numbers of substances making it to market launch decline, and their successes do not necessarily compensate for the increased investment.

It has been clearly shown that the problem lies in the clinical segment, i.e., when the promise of (mostly) animal model-based final selection of a compound to advance to the clinics is challenged. In fact, failure rates in the clinical phase of development now reach 95% (Arrowsmith, 2012). Obviously, recent biomedical research breakthroughs have not improved our ability to identify successful candidates. The main causes of failure in the clinic include safety problems (about 20%) and lack of effectiveness (about 40%), both predicted by a series of animal models before entering the most costly phase of drug development. The inability to predict these failures before human testing or early in clinical trials dramatically escalates costs. Overall, biopharmaceuticals appear to have the higher success rate (all indications) of 30.2% (Gilbert et al., 2003), exactly those which can often not be tested in animals as they are human-specific in their effect.

This illustrates the limited predictivity of our predominant tool – the animal model. We have made these points in more detail earlier (Hartung, 2008b, 2009b, 2013). Pertinent summaries are available from different sources (Pound et al., 2004; Olson et al., 2000; Hartung, 2008b) showing reasons for differences between animal studies and human trials as recently summarised (Hartung, 2013). The problem became very evident in case of medical countermeasures for biological and chemical terrorism and warfare (Hartung and Zurlo, 2012). The author had the privilege to serve on the National

Academy of Sciences panel on animal models for countermeasures to bioterrorism. The problem of developing and stockpiling drugs for the event of biological/chemical terrorism or warfare is that (fortunately) there are no patients to test on. The question to the panel, therefore, was: how do we substitute (in conformance with the animal rules of the FDA) with suitable animal models? In a nutshell, our answer is: "There is no such thing as a sufficiently predictive animal model substitute for clinical trials" (NRC, 2011). Any drug company would long to have such models for drug development, as the bulk of development costs are part of the clinical phase only. For countermeasures we have even more difficulties: unknown pathophysiology, limitations to experiment in biosafety facilities, disease agents potentially designed to resist interventions, and mostly peracute diseases. Therefore, an important part of the committee's discussions dealt with the attrition (failure) rate of drugs entering clinical trials (see above), which does not encourage the use of animal models as a substitute for clinical trials. Similarly, a recent paper by Seok et al. (2013) showed the lack of correspondence of mouse and human responses in sepsis, probably the clinical condition closest to biological warfare and terrorism as discussed earlier (Leist and Hartung, 2013).

The quoted data from Arrowsmith would suggest that toxic side effects contribute to 20% of attrition (each) in phases II and III. We likely need to add to that percentage for side-effects noted in phase I, i.e., first in humans, and post-market adverse reactions. Unexpected side effects lead to withdrawals – Wikipedia lists 59 drugs withdrawn from the market over the last three decades[2], representing roughly 10% of new drugs entering the market. This does not even include drugs where indications had to be limited because of problems. Thus, an overall figure of 30%–40% seems realistic.[3] The hallmark paper by Olson et al. (2000) gives us some idea of this and the retrospective value of animal models in identifying such problems. Rats and mice, for example, predicted together, only 43% of clinical toxicities of candidate drugs observed later. In toxicology, we have seen that different laboratory species exposed to the same high doses predict each other no better than 60% (Basketter et al., 2012) – and there is no reason to assume that any of them predict human toxicity better at low doses. We lack drug efficacy models for systematically comparing outcomes in different strains or species of laboratory animals. It is unlikely that results on their predictivity are much better.

2.4 Precautionary approaches from drug development adapted to other areas

Most of our methods were introduced for the safety evaluation of drugs under development. These are substances meant to be bioavailable and have biological effects. Typically, there are also a number of structural variants from which to choose the lead compound in case a toxicological problem develops. At this stage in drug development, most importantly, there is neither experience with human exposure nor a market value (other than the development costs). So, does a toxicology designed for pharmaceutical substances under development, and used (to some extent) for new, low-production volume chemicals over the last decades, suit us for the assessment of existing high-production volume chemicals? Not without, at minimum, an assessment of a suitability of each and every standard test. We need to determine how cautious we want to be. The hassle involved in restricting the use of, or substituting for substances with, complex use scenarios may be considerable and should be undertaken only if well warranted. We also need to ask whether the methods are applicable to the type of substances. We may think that we have a lot of experience due to the Dangerous Substance Directive, which was in effect from 1981 until it was replaced by REACH, or the parallel Toxic Substances Control Act in the USA. For many tests, despite being prescribed for three decades for new chemicals, we have, in fact, minimal experience with industrial chemicals because they were not triggered for the relatively small production volumes of new chemicals. We found 14 cancer bioassays and 46 two-generation studies in 28 years for approximately 4,500 notifications in Europe. It is easy to accept major testing demands for high-production volume chemicals if they are never applied. Now, 35 years later, it is obviously difficult to complain about their implementation in the first place. Adapting to methods with acceptable false-positive rates or the introduction of mechanisms to sort out false-positive results should be a goal of programs such as REACH. Today, we consider the guideline tests as 'definitive' which can result in up to 20 times more false- than real-positives as argued earlier for reproductive toxicity testing (Hartung, 2009b).

2.5 Animal use

Concerns about animal experimentation and the killing of animals have a long history. Even the ancient Greeks discussed whether we should kill animals. In Germany in the 1920s, there were 700 animal welfare associations. However, it was not until 1959 that Russell and Burch (1959) in England developed what they called "the principles of humane experimental technique". They referred to these principles as the '3Rs,' i.e., *Replacement*, *Reduction*, and *Refinement*, and are defined thusly:

1 one must substitute animals with non-sentient test systems (Replacement) when an alternative exists

2 one must reduce the number of animals used wherever possible if the same result can be obtained with fewer animals (Reduction)

3 one must avoid unnecessary suffering and distress by using, for example, analgesics or working under narcosis (Refinement).

Any suggestion of Replacement was actually utopian 60 years ago as, at that time, cell culture and computer programs were in their infancy and few scientists could imagine that such methods might lead to predictive tools. Over the last few decades, however, industry, science, and politics have demonstrated a commitment to the 3Rs that has made it the compromise with those who would prefer to see animal experiments end today rather than tomorrow. The 3Rs became the basis for a credible investment in overcoming reliance animal testing. In fact, animal experiments decreased until the turn of the century by an estimated two-thirds from their peak in the mid-1970s. Since then, however, numbers have been increasing, due largely to the new techniques for manipulating individual genes in mice, which have become very popular scientific models.

In the meantime, we have a number of 3Rs approaches that have come to fruition. The LD_{50} test, for example, which determines the lethal dose of a chemical that kills 50% of treated rats, has been used since the 1920s. Until 1989, 150 animals per substance were used in this test (10 female and 10 male at 7 dosages each, plus one untreated control group of 10 animals). This resulted in an enormous number of animals being used, since nearly every substance going to market was tested. Both the protection of workers and safety measures for the transport of substances also were determined on this basis. In

1989, after an analysis of test data, a revision of guidance took place at the Organization for Economic Co-operation and Development (OECD). The OECD, now with 34 industrialised countries, achieved agreement to drastically reduce the number of animals used in the LD_{50}. Since then, groups of five animals of one gender have been used, reducing the number of rats from 150 to 45 per substance. In the 1990s, a further step was taken. The idea was simple: Why should all animals be treated simultaneously? When starting with just one dose, a higher dose can be tested next if the animals survive. If the animals die, the dose has to be lowered. At the same time, it was shown that groups of three rats suffice. Consequently, three methods were accepted internationally in 2001.[4] On average, these tests use only 8 to 12 animals. From 150 to 45 to just 8–12 animals – an enormous reduction indeed. In addition, one of these methods introduced the notion that the animal does not have to await death but rather can be euthanised humanely when it shows signs it will not survive or will be severely damaged. This is an example of refinement – the second R – for the amelioration of pain and distress in animal experiments. Another classical example is testing for skin allergy. Traditionally, this has been done with guinea pigs. The local lymph node assay (LLNA) represents both a reduction and a refinement alternative in mice, as it uses fewer animals, involves a shorter treatment period, and ends the experiment at the stage of lymph node swelling instead of waiting for skin lesions to occur.

Increasingly, animal tests in toxicology can be fully substituted – i.e., the third R, replacement. As an example, human skin obtained from surgical procedures can be grown in the laboratory. A small tissue sample can produce several square metres of skin. These technologies were developed originally for skin transplantation after burn injury, but the idea quickly arose that this tissue could also be used for testing chemicals. In fact, it was possible to demonstrate that artificial human skin is as suitable as rabbit skin for testing skin corrosion or chemical irritation. The respective international test guidelines have been agreed upon. This was not only a milestone for the cosmetic industry (Hartung, 2008b), but also proof-of-principle that international consensus can be achieved regarding the replacement of an animal test with an animal-free method (Hartung and Daston, 2009).

We have to be clear, however, that these first successes do not mean that we can replace all animal testing. These were the low-hanging fruits of acute and topical effects. However, strategies to replace the more challenging systemic and repeated-dose toxicity tests are emerging (Basketter et al., 2012; Leist et al., 2014). Increasing public concern about the use of animals in research affects scientists, whether they share the concerns or not, as legislation increasingly creates more restrictions [as in the most recent laboratory animal welfare legislation in Europe (Hartung, 2010a)]. These forces are as much a driver for change as scientific progress.

2.6 New products not suitable for traditional tests

Toxicology has largely developed around the safety assessment of drugs and environmental chemicals. A number of product categories do not always fit this (mainly) animal-based toolbox:

1. *Biologicals* (biopharmaceutical products), i.e., therapeutic substances, such as a vaccine or drug, derived from biological sources, especially if they are human proteins such as antibodies or cytokines. More than 50% of new drugs fall into this

category. Here, species-specificity poses problems. Target-related (side-)effects (excess pharmacology), for example, frequently cannot be seen in animals. Examples include variants in the molecule itself between species, in binding sites, or induced responses. As these are foreign materials, such as proteins different from those of the laboratory animal species, animals will often develop (neutralising) antibodies and can even develop anaphylactic reactions if chronically treated. While traditional toxicological endpoints are of similar relevance, the immunomodulatory properties of these entities require additional focus (Hartung and Corsini, 2013). Often, production is also more difficult to standardise, requiring batch release tests (as is well known for traditional vaccines).

2 *Cell therapies*. Again, we have to deal with problems of species-specificity and immunological reactions when repeatedly applied.

3 *Genetically, modified and functional foods (nutraceuticals)*. Often these test animals are given inadequate feed, when large parts of their total intake are the test material. This is problematic as it can have effects unrelated to the material in question and lacking human relevance. In general, genetic changes in food and feed are not expected to result in health effects, which has not hindered extensive animal testing requirements, sometimes giving the impression that these serve mainly as a roadblock to entering the market (Hartung and Koeter, 2008; EFSA GMO Panel Working Group on Animal Feeding Trials, 2008).

4 *Medical countermeasures to biological and chemical terrorism and warfare agents*. In this instance, the difficultly lies more in efficacy testing than safety assessment (Hartung and Zurlo, 2012).

5 *Medical devices* are typically solid, often made from a variety of materials, and can have mechanical, often long-lasting interactions with the body. Many reactions occur on the surface, which is difficult to mimic except in settings similar to their clinical use. Surface contaminations and leaching are typical problems, but eluates and rinsing solutions cannot reflect the potentially high local concentrations. Again, standardisation of the product can be problematic, requiring batch-release testing.

6 *Nanoparticles*. Nanoparticle (NP) toxicology is a rapidly growing concern (Hartung, 2010c; Hartung and Sabbioni, 2011), driven by the dramatic increase in industrial uses of NP and fuelled by public debate. Increasing funding and studies inevitably will result in reports of toxicological effects of NP – both publication bias for positive findings and the multiple testing fallacy (if 20 experiments or endpoints are analysed with $p = 0.05$ for significance, one should be false-positive) will come to bear here. NP can be seen as extreme forms of medical devices – solid materials with enormous surfaces. This can affect reactivity, kinetics, and produce new (unexpected) biological properties. Experiences from particulate matter (PM) suggest that lung and cancer effects should be considered for engineered nanomaterials, and these endpoints are not the strength of the traditional testing battery. Furthermore, possible associations of PM with asthma and atherosclerosis suggest a need for new endpoints to be covered (see below) by toxicologists. So far, it appears that the problem of nanotoxicology is mainly a kinetic one – some safety factors could help to account for differences in ADME, but we need to keep in mind that the enhanced bioavailability of NP on body, organ, and cellular levels might result in thresholds of

toxicity being overstepped – in which case, a change in hazards suddenly becomes relevant.

These testing needs can be seen as door openers for new methods, which will hopefully enrich our toolbox of traditional safety assessments. A prominent example is the current developments of *human-on-chip* approaches prompted by the needs of medical countermeasures (Hartung and Zurlo, 2012).

2.7 New hazards not covered

Possible examples of continuously increasing health problems include atherosclerosis, male infertility and other possible manifestations of endocrine disruption, autism and other developmental behavioural disorders (Smirnova et al., 2014), immunotoxicity (Hartung and Corsini, 2013), childhood asthma and other obstructive lung diseases, obesity, and diabetes. This is not to suggest that these health problems are caused by environmental chemicals, but that their unclear etiology and, in most cases, increasing numbers, suggest the need for testing. As an example, it is worth noting that air pollution involving natural NP leads mainly to excess deaths associated with cardiovascular illness (Seaton and Donaldson, 2005), a hazard not generally addressed in toxicology. Arteriosclerosis, in fact, is very difficult to induce in animals. Determination of the pulmonary and systemic inflammatory hazards typically seen with NP (Kipen and Laskin, 2005) is not among the strengths of the toxicological toolbox.

The struggle to develop a toolbox for endocrine disruptors is a good example of potential new hazards that are changing toxicology (Dietrich, 2010; Juberg et al., 2013). This has led to new technologies, too, as a number of *in vitro* tests and more recent high-throughput testing strategies have been introduced in this process. Many of the unexplored health and environmental effects may introduce new technologies as well – not just adding to the existing technologies, but helping to improve, backload, and replace them.

Diseases of environmental origin and new exposures are believed to be complementary, and exposures to one or more of the new hazards might produce new outcomes and new diseases (Genuis 2012). Many diseases (e.g., Alzheimer’s disease, asthma, autism, cancer, cardiovascular problems, diabetes, mental illnesses, multiple sclerosis, obesity, and Parkinson’s disease, to name a few) are regularly shown to have both environmental and genetic causes and emphasis on one or the other typically comes in waves. The sequencing of the human genome created a ‘hyped’ incentive to explain disease more on the basis of genetic repertoire, but the subsequent genome-wide association studies resulted in very few (and not very strong) associations – a major disappointment (Janssens and van Duijn, 2008). Thus, environmental theories of pathogenesis are on the upswing again. And, with the advent of epigenetics, we may now, for the first time, have a common explanation for genetic mechanisms in the manifestation of environmental exposures. Here, it should suffice to say that the need to understand the pathogenesis of many diseases in order to develop prevention and treatment solutions is a continuous driving force for new toxicological approaches. Current approaches have little to offer, as very few diseases have been clearly attributed to environmental causes based on experimental laboratory animal studies. On a mechanistic level, we might actually determine how environmental toxicants induce and aggravate disease processes, further fuelling the need for the novel mechanistic approaches to toxicology.

2.8 Mixtures of toxicants not addressed

Mixture toxicology is a holy grail of toxicological science. Why so? First of all, because it represents actual exposures. Secondly, because of the enormous number of combinations of even just two substances – which is endless. We can vary doses and timing in any possible way and each of these dimensions does not need to be static. For example, we might change doses or timing of exposures during the course of an experiment. This would actually reflect our exposure to environmental chemicals most closely. And there is no reason to limit our tests to two substances. The situation seems to be more defined for fixed mixtures (formulations) such as end products, where we have considerable testing experience (though typically either the end product or its ingredients – rarely both – are tested for regulatory purposes). Here at least, species differences vis-à-vis humans come in, amplified by the number of biologically active components and their differing kinetics (including metabolism). Since many of these differences are interlinked, we have to expect amplified diversity for mixtures between species. The problem is multifaceted and needs to be addressed as such – ranging from questions about end-product versus ingredient testing, minimal versus significant product mixture changes requiring re-testing, to the endless patterns of individual combined exposures with unknown toxicological impacts. The only obvious conclusion is that there is a lot to test – far more than we can afford in traditional animal tests. At costs of tens to hundreds of thousands of dollars for testing a single substance in a few fixed concentrations, traditional regulatory animal tests are not suitable. The new high-throughput approaches at least promise to allow addressing the principles of interactions. If at all, high-throughput and, to some extent, high-content assessments, can help address mixtures, but those are more 'basic science' approaches than appropriate for regulatory purposes.

2.9 Individual susceptibilities and vulnerable subpopulations are not covered

If mixture toxicology is the holy grail, human individual susceptibilities are the 'mission impossible' for traditional approaches. Even if genetically diverse animal populations can be used, they provide little information about the combination of genetics, lifestyle, and past exposures (epigenetics) that form the individual. We might be better off studying subpopulations, as some aspects of susceptibility might be shared – for example, by young or elderly between humans and animal species – though many aspects differ considerably and older, purpose-bred laboratory animals are prohibitively expensive. The best approach would be studying the pathways of toxicity (PoT) and combining this information with polymorphisms and other individual features (such as kinetics) for a systems toxicology approach (Hartung et al., 2012), thus modelling inter-individual differences.

2.10 Poor basic research and publication standards

While the previous challenges were centred on toxicology, with the exception of animal study predictivity, the last challenge is one shared by most life sciences. A more self-critical view is necessary not only in regulatory science, where it is prompted by public health and economical consequences, but also in basic toxicological research. The reader is invited, however, to translate these thoughts to other areas of the life sciences.

We have recently addressed the limitations of pre-clinical research (Hartung, 2013). While there was an intense discussion about clinical study standards and their reporting – even prominent criticism that reached lay audiences – toxicology and basic research are only beginning to be included in the discussion. Clinical studies, in fact, have extremely high quality standards. They are mostly randomised, double-blind, and placebo-controlled, and usually multi-centric. They require ethical review, follow good clinical practice guidelines, and are carried out by skilled professionals. In recent years, the push to publish and register clinical trials has increased dramatically. Clinical medicine also gave rise to evidence-based medicine (EBM), which we have praised as an objective, transparent, and conscientious way to condense information for a given controversial question (Hoffmann and Hartung, 2006; Hartung, 2009a, 2010b). Altogether, these are attributes difficult to match in other fields.

We might say that clinical research is good at acknowledging its biases and overestimating its successes. Put simply, the clinical pipeline, despite enormous financial pressures, has very sophisticated tools to promote high-quality science. Compared to clinical studies, toxicology has some advantages and some disadvantages when it comes to quality. First, there are internationally harmonised protocols [especially International Conference on Harmonisation (ICH) and OECD] and Good Laboratory practice (GLP) guidances for quality assurance. Our methods, however, are outdated – most were introduced before 1970, and were systematically rendered precautionary/oversensitive (for example, by the use of extremely high doses). The mechanistic thinking of modern toxicology can be seen as (in the Dutch idiom) 'mustard after the meal': that is, mostly for arguing why the animal findings are not relevant to humans. What is most evident when comparing approaches is that clinical studies have one endpoint, good statistics, and hundreds to thousands of treated individuals with relevant exposures. Toxicology does everything just the opposite – group sizes of identical twins (inbred strains) are minimal, and we study a large array of endpoints (often at 'maximum tolerated doses') without proper statistics. The only reason is feasibility: we limit group sizes to what is affordable (and certainly from the perspective of animal use, desirable). But these compromises combine in the end and limit the relevance of prediction if, for example, we do not have the statistical power to control for multiple testing.

Scientific relevance demands reproducibility. Two recent publications by authors from major pharmaceutical companies can be seen as an epiphany: both Amgen (Begley and Ellis, 2012) and Bayer HealthCare (Prinz et al., 2011) showed that they essentially could not reproduce the key findings of many studies that had prompted drug development. How is this possible? Basic researchers seem to be even more naïve in the interpretation of results than clinical researchers. In a comparison of 108 studies (Lumbreras et al., 2009), laboratory scientists were 19 times more likely to over-interpret the clinical utility of molecular diagnostic tests than clinical scientists. Basic research (at least in academia, where it is the source of most such papers) is conducted mostly un-blinded in a single laboratory and executed by students learning on the job, normally without any formal quality assurance measures. Limited replicates due to limited resources and time, as well as pressure to publish, lead to publications that do not always withstand replication. Insufficient documentation aggravates the situation.

The weaknesses in quality and reporting of animal studies, in particular, have been demonstrated (MacCallum, 2010; Macleod and van der Worp, 2010; Kilkenny et al., 2010; van der Worp and Macleod, 2011), further undermining their value. The availability of the ARRIVE guideline (Kilkenny et al., 2010) and the Gold Standard Publication Checklist (GSPC) to improve the quality of animal studies and their reporting (Hooijmans et al., 2010) facilitate the evaluation and standardisation of publications on animal studies. Randomisation and blinding are rarely reported, which can have important implications, as it has been shown that animal experiments carried out without either is five times more likely to report a positive treatment effect (Bebarta et al., 2003). Baker et al. (2012) recently gave an illustration of poor reporting on animal experiments showing that out of 180 papers on multiple sclerosis only 40% used appropriate statistics, as did only 4% of neuroimmunological studies published in two years in *Nature*, *Science* and *Cell* (Baker et al., 2012). Poor statistics are a more widespread problem than outsiders might believe. Awareness is a little better in clinical research (Andersen, 1990; Altman, 1994, 2002), but too often we reviewers or readers see papers without statistics or with inappropriate statistics (such as the promiscuous use of t-tests where not justified). Some common mistakes were illustrated in Festing (2003), Lang (2004) and Altman (1998). Altman (1998) summarised 13 previous analyses of the quality of statistics in medical journals. The 1,667 papers analysed show that only about 37% have acceptable statistics, and the trend has not improved over time. When asking why many scientific papers are wrong, even if statistics are correctly applied, we also have to consider that a study usually does not depend on a single experiment. Instead, we report on a number of experiments that, when taken together, make the case. Even if we achieve a significance level of 95% in each given experiment, when combined, the probability of an error increases steadily. When seven experiments lead to a joint conclusion, each with 95% probability, the overall conclusion is only 70% certain. But, the purpose of this article is not a review of statistics and statistical practice. Rather, it illustrates yet another contributor to the non-reproducibility of results. We might summarise with a quote from Lang: "He uses statistics as a drunken man uses lamp-posts – for support rather than illumination."

The problems are by no means less significant in cell culture work. Our attempts to establish Good Cell Culture Practice (GCCP) (Coecke et al., 2005) and publication guidance for *in vitro* studies (Leist et al., 2010) desperately await broader implementation (see below). Earlier, we have discussed the shortcomings of typical cell cultures (Hartung, 2007, 2013). These articles built upon experiences gained from the validation of *in vitro* systems and in the course of developing the GCCP guidance (Coecke et al., 2005). Noteworthy, this activity has been recently reactivated to develop GCCP 2.0 to include stem cells and organotypic cultures. Cell cultures are prone to artefacts (Hartung, 2007a) and far too many artificially chosen and difficult-to-control conditions influence our experiments. Quality assurance is the gift from alternative methods to the life sciences. This blunt statement might be challenged by those involved with GLP or ISO quality assurance. However, while GLP (at least originally) addressed only regulatory *in vivo* studies, and ISO guidance is not specific for life science tools, neither addresses the key issue – the relevance of a test. This is the truly unique contribution of validation, which is far too infrequently applied in other settings.

We do not obtain *in-vivo*-like differentiation because we often start with tumour cells (tens of thousands of mutations, loss and duplications of chromosomes), over-passaging with selection of subpopulations, non-physiologic culture conditions (minimal cell contact, low cell density, no polarisation, limited oxygen supply, non-homeostatic media exchange, temperature and electrolyte concentrations reflective of humans not rodents), forcing growth (fetal calf serum, growth factors), no demand on cell functions due to over-pampering, and no *in vitro* kinetics giving consideration to the fate of test substances in the culture and lack of cell type interactions. For most aspects there are technical solutions, but few are applied, and if they are applied it is done in isolation, solving some – but not all – of the problems. On top of this, we lack quality control. If we accept the following estimates, it is likely that only 60% of studies use the intended cells without mycoplasma infection. Misidentified cells are a threat to all *in vitro* work. The most impressive are HeLa cells. Since 1967, cell line contaminations have been evident, i.e., another cell type was accidentally introduced into a culture and slowly took over. The most promiscuous so far are HeLa cells, actually the first human tumour cell line. The line was derived from cervical cancer cells taken in 1951 from Henrietta Lacks, a patient at Johns Hopkins. The cells have contributed to more than 60,000 research papers and the development of a polio vaccine in the 1950s (Skloot, 2010). Recently, the HeLa genome has been sequenced (Landry et al., 2013). It is most interesting to see the genetic make-up of the cells as summarised by Ewen Callawa in *Nature*:[5]

> "HeLa cells contain one extra version of most chromosomes, with up to five copies of some. Many genes were duplicated even more extensively, with four, five, or six copies sometimes present, instead of the usual two. Furthermore, large segments of chromosome 11 and several other chromosomes were reshuffled like a deck of cards, drastically altering the arrangement of the genes."

Do we really expect such a cell monster to show normal physiology? The cell line was found to be remarkably durable and prolific, as illustrated by its contamination of many other cell lines. It is assumed that, even today, 10%–20% of cell lines are actually HeLa cells and, in total, 18%–36% of all cell lines are wrongly identified. Even over the last decade, studies analysing the problem of inauthenticity in cell banks range from 15%–18% (Hughes et al., 2007) and a very useful list of such mistaken cell lines is available.[6] The problem has been raised several times (Macleod et al., 1999; Stacey, 2000; Buehring et al., 2004; Rojas et al., 2008; Dirks et al., 2010). A study (Buehring et al., 2004) from 2004 showed that HeLa contaminants were used unknowingly by 9% of survey respondents, a likely underestimation of the problem, and only about a third of respondents were testing their lines for cell identity. It is a scandal that a large percentage of *in vitro* research is done on cells other than the supposed ones and, thus, misinterpreted.

Another type of contamination that is astonishingly frequent and has a serious impact on *in vitro* results is microbial infection, especially with mycoplasma (Langdon, 2003). Screening by the FDA for more than three decades shows that, of 20,000 cell cultures examined, more than 3,000 (15%) were contaminated with mycoplasma (Rottem and Barile, 1993). Studies in Japan and Argentina reported mycoplasma contamination rates of 80% and 65%, respectively (Rottem and Barile, 1993). An analysis by the German Collection of Microorganisms and Cell Cultures (DSMZ) of 440 leukemia-lymphoma cell lines showed that 28% were mycoplasma positive (Drexler and Uphoff, 2002). Laboratory personnel are the main sources of *M. orale*, *M. fermentans*, and *M. hominis*. These species of mycoplasmas account for more than half of all mycoplasma infections in cell cultures and physiologically are found in the human oropharyngeal tract

(Nikfarjam and Farzaneh, 2012). *M. arginini* and *A. laidlawii* are two other mycoplasmas contaminating cell cultures that originate from fetal bovine serum (FBS) or newborn bovine serum (NBS). Trypsin solutions provided by swine are a major source of *M. hyorhinis*. It is important to understand that the complete lack of a bacterial cell wall of mycoplasma implies resistance against penicillin (Bruchmüller et al., 2006), and they even pass 0.2 μm sterility filters, especially at higher pressure rates (Hay et al., 1989). Mycoplasma can have diverse negative effects on cell cultures, and it is extremely difficult to eradicate this intracellular infection (Drexler and Uphoff, 2002; Nikfarjam and Farzaneh, 2012). While there is good understanding in the respective fields of biotechnology, this is much less the case in basic research, and mycoplasma testing is neither internationally harmonised with validated methods nor common practice in all laboratories on a regular basis. The recent production of reference materials (Dabrazhynetskaya et al., 2011) offers hope for the respective validation attempts. The problem lies in the fact that there are at least 20 different species found in cell culture, though 5 of them appear to be responsible for 95% of the cases (Bruchmüller et al., 2006). For a comparison of the different mycoplasma detection platforms see Lawrence et al. (2010) and Young et al. (2010).

The documentation practices in laboratories and publications are often sub-par. There is some guidance available (GLP has been increasingly adapted; for GCCP see below) but it is infrequently applied. The more recent mushrooming of cell culture protocol collections is an important step, but adherence is still uncommon and deviations are unclear in publications. We tend to toy around with the models until they work for us, and too often *only* for us. Such standardisation forms the basis for formal validation, as developed by ECVAM, adapted and expanded by ICCVAM and other validation bodies, and, finally, internationally harmonised by OECD (2005). Validation is the independent assessment of the scientific basis, the reproducibility, and the predictive capacity of a test. It was redefined in 2004 in the Modular Approach (Hartung et al., 2004) but needs to be seen as a continuous adaptation of the process to practical needs and a case-by-case assessment of what is feasible (Hartung, 2007b, Leist et al., 2012). The most important changes to the Modular Approach were the introduction of an applicability domain (borrowing the concept from QSAR), the use of existing data (retrospective validation), and the independence of reproducibility and relevance assessment, which allows for leaner study designs and performance standards for similar tests to be considered equivalent to a validated one. The framework of EBM is increasingly being translated to toxicology (Hoffmann and Hartung, 2006), and it recently led to the creation of the Evidence-based Toxicology Collaboration (EBTC)[7] (Zurlo, 2011; Stephens et al., 2013; Hoffmann et al., 2014).

The advent of human embryonic and induced pluripotent stem cells appears to be something of a game-changer. First, it promises to overcome the problems of availability of human primary cells, though a variety of commercial providers make almost all relevant human cells available in reasonable quality (but at costs that are challenging for academia). It is important to note, however, that we do not yet have protocols to achieve full differentiation of any cell type from stem cells. This is probably only a matter of time, but many of the non-physiological conditions from traditional cell cultures contribute here. Stem cells have been praised for their genetic stability, which appears to be better than for other cell lines, but we have increasingly discovered their limitations in that respect, too (Mitalipova et al., 2005; Lund et al., 2012; Steinemann et al., 2013). The limitations experienced first are costs of culture and slow growth – many protocols require months, and labour, media, and supplement costs add up. The risk of infection increases unavoidably. We still do not obtain pure cultures and often require cell sorting, which, however, implies detachment of cells with the respective disruption of culture conditions and physiology.

Science is increasingly aware of the shortcomings of its approaches. Ioannidis (2005b, 2005a) has stirred controversy with papers entitled 'Why Most Published Research Findings Are False' (excerpt: "for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias") and "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research". As early as 1994, Altman (1994) wrote of "The scandal of poor medical research". This does not even address the contribution of fraud (Fang et al., 2012). These early warnings have now been substantiated with the unsuccessful attempts by industry to reproduce important basic research (see earlier). Drummond Remmie wrote:

> "Despite this system, anyone who reads journals widely and critically is forced to realize that there are scarcely any bars to eventual publication. There seems to be no study too fragmented, no hypothesis too trivial, no literature too biased or too egotistical, no design too warped, no methodology too bungled, no presentation of results too inaccurate, too obscure, too contradictory, no analysis too self serving, no argument too trifling or too unjustified, and no grammar and syntax too offensive for a paper to end up in print. The function of peer review, then, may be to help decide not whether but where papers are published."

The situation is not very different if this is *in vitro* or *in vivo* work, (and the two are often combined). Similar things can be said about *in silico* work (Hartung and Hoffmann, 2009), which is not only limited by the *in vitro* and *in vivo* data it is based on (trash in, trash out), but has inherent problems of lack of data accuration and overfitting. "Torture numbers, and they'll confess to anything" (Gregg Easterbrook). One difference is that *in vitro* approaches have developed the principles of validation. There is no field more self-critical than the area of alternative methods, where we spent one-half to one million dollars and, on average, ten years, to validate a particular method. Basic research could learn from this in its quest to establish the reproducibility and relevance of our methods. We are not calling for GLP for academia, but rather for the spirit of GLP to be embraced.

Clinical research should be a role model for basic research and regulatory sciences, especially evidence-based approaches, documentation, and quality assurance. Publish less, but of better quality, or as Altman (1994) put it: "We need less research, better research, and research done for the right reasons."

3 Conclusions

This article identified ten major reasons why toxicology has to change.

1 because we treat different substances very differently depending on product category or geographic region, while we treat substances the same, which have very different uses

2 because throughput and costs of current testing cannot satisfy the societal testing needs and do not allow early assessments in product development and design guidance (Green Toxicology)

3 because predictivity for humans (and even of different animal species) is limited

4 because precautionary approaches from drug development adapted to other areas produce too many false-positives

5 because animal use is seen more and more critically by a general public, thus affecting legislation

6 because new products (biological, cell therapies, genetically modified and functional food (nutraceuticals), medical countermeasures to biological and chemical terrorism and warfare agents, as well as medical devices and NPs) are not always suitable for traditional tests

7 because new and emerging hazards (e.g., endocrine effects, childhood effects such as asthma and behavioural issues, obesity, and cardiovascular effects) are not adequately covered

8 because mixture effects of toxicants cannot be adequately addressed

9 because individual susceptibilities and vulnerable subpopulations cannot yet be covered

10 because science in general has to raise research and publication standards.

It is perhaps too easy to just criticise, but the challenge to the author was to summarise why current approaches require change. While this article focuses mostly on toxicology, we have attempted to extend some critical observations to research in general. This will, first of all, show that toxicology's problems are not different, and that the field is perhaps even advanced with regard to internationally harmonised methods and quality assurance. New approaches are emerging, under the banner of 'Toxicology for the 21st Century,' that rely on molecular pathways of human toxicity. These include activities for implementing the 2007 NRC report on *Toxicity Testing for the 21st Century: A Vision and a Strategy* such as the Hamner case study approach (Andersen et al., 2011), the EPAs ToxCast program (http://www.epa.gov/ncct/toxcast/), as well as the US multi-agency alliance Tox21 (http://www.ncats.nih.gov/research/reengineering/tox21/tox21.html), which test thousands of substances in high-throughput screening assays and make the highly quality-controlled data publicly available, or the NIH Human Toxome project (Hartung and McBride, 2011; Bouhifd et al., 2015), which is beginning to map PoT in a systematic manner. On the European side, a number of projects, most prominently SEURAT-1 (Gocht et al., 2015), aim to develop the use of novel technologies for a variety of toxicological applications; the differences in approach have been characterised earlier (Hartung, 2010e). There has also been significant development of biomarkers and there is the need for the continued collaboration of toxicologist with epidemiologist in the development of molecular epidemiology.

The new approaches will have to show whether they can more adequately address these challenges. It all starts, however, with being aware of these challenges and tackling them. It requires our willingness to change practices and not waste time defending traditional approaches. Futurist Alvin Toffler wrote: "The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn." Progress in toxicology, if stilted now, may worsen because the increasing educational specialisation in the field at the expense of the toxicologist who in training and experience previously developed a sufficiently broad knowledge to identify and draw upon the expertise of specialists, and also to identify its ‘big picture’ needs. We still need the current approaches – these methods have helped to make the world a safer place. There is more than enough reason, however, to advocate for substantial change.

References

Altman, D.G. (1994) ‘The scandal of poor medical research’, *BMJ: British Medical Journal*, Vol. 308, No. 6924, pp.283–284.

Altman, D.G. (1998) ‘Statistical reviewing for medical journals’, *Statistics in Medicine*, Vol. 17, No. 23, pp.2661–2674.

Altman, D.G. (2002) ‘Poor-quality medical research – what can journals do?’, *JAMA: The Journal of the American Medical Association*, Vol. 287, No. 21, pp.2765–2767.

Ames, B.N. and Gold, L.S. (2000) ‘Paracelsus to parascience: the environmental cancer distraction’, *Mutat. Res.*, Vol. 447, No. 1, pp.3–13.

Andersen, B. (1990) *Methodological errors in medical research: An Incomplete Catalogue*, 288pp, Blackwell Science Ltd., Chicago.

Andersen, M.E., Clewell III, H.J., Carmichael, P.L. and Boekelheide, A.K. (2011) ‘Can case study approaches speed implementation of the NRC report: ‘Toxicity Testing in the 21st Century: A Vision and a Strategy?’’, *ALTEX*, Vol. 28, No. 3, pp.1–8.

Arrowsmith, J.J. (2012) ‘A decade of change’, *Nature Reviews Drug Discovery*, Vol. 11, No. 1, pp.17–18.

Baker, D.D., Lidster, K.K., Sottomayor, A.A. and Amor, S.S. (2012) ‘Reproducibility: research-reporting standards fall short’, *Nature*, Vol. 492, No. 7427, pp.41–41.

Ball, N., Cronin, M.T.D., Shen, J., Blackburn, K., Booth, E.D., Bouhifd, M., Donley, E., Egnash, L., Hastings, C., Juberg, D.R., Kleensang, A., Kleinstreuer, N., Kroese, D., Lee, A.C., Luechtefeld, T., Maertens, A., Marty, S., Naciff, J.M., Palmer, J., Pamies, D., Penman, M., Richarz, A-N., Russo, D.P., Stuard, S.B., Patlewicz, G., van Ravenzwaay, B., Wu, S., Zhu, H., and Hartung, T. (2016) ‘Toward good read-across practice (GRAP) guidance’, *ALTEX*, Vol. 33, No. 2, pp.149–166 [online] http://doi.org/10.14573/altex.1601251.

Basketter, D.A., Clewell, H. and Kimber, I. et al. (2012) ‘A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing – t4 report’, *ALTEX*, Vol. 29, No. 1, pp.3–91.

Bebarta, V., Luyten, D. and Heard, K. (2003) ‘Emergency medicine animal research: does use of randomization and blinding affect the results?’, *Academic Emergency Medicine*, Vol. 10, No. 6, pp.684–687.

Begley, C.G. and Ellis, L.M. (2012) ‘Drug development: raise standards for preclinical cancer research’, *Nature*, Vol. 483, No. 7391, pp.531–533.

Bottini, A.A. and Hartung, T. (2009) ‘Food for thought... on the economics of animal testing’, *ALTEX*, Vol. 26, No. 1, pp.3–16.

Bottini, A.A. and Hartung, T. (2010) ‘The economics of animal testing’, *ALTEX*, Special Issue, Vol. 27, No. 1, pp.67–77.

Bottini, A.A., Amcoff, P. and Hartung, T. (2007) ‘Food for thought ... on globalisation of alternative methods’, *ALTEX*, Vol. 24, No. 4, pp.255–269.

Bouhifd, M., Andersen, M.E., Baghdikian, C., Boekelheide, K., Crofton, K.M., Fornace Jr., A.J., Kleensang, A., Li, H., Livi, C.B., Maertens, A., McMullen, P.D., Rosenberg, M., Thomas, R., Vantangoli, M., Yager, J.D., Zhao, L. and Hartung, T. (2015) ‘The Human Toxome project’, *ALTEX 2015*, Vol. 32, pp.112–124.

Briel, M., Müller, K.F. and Meerpohl, J.J. et al. (2013) ‘Publication bias in animal research: a systematic review protocol’, *Systematic Reviews*, Vol. 2, pp.23–23.

Bruchmüller, I., Pirkl, E. and Herrmann, R. et al. (2006) 'Introduction of a validation concept for a PCR-based Mycoplasma detection assay', *Cytotherapy*, Vol. 8, No. 1, pp.62–69.

Buehring, G.C., Eby, E.A. and Eby, M.J. (2004) 'Cell line cross-contamination: how aware are mammalian cell culturists of the problem and how to monitor it?', *In Vitro Cellular & Developmental Biology – Animal*, Vol. 40, No. 7, pp.211–215.

Coecke, S., Balls, M. and Bowe, G. et al. (2005) *Guidance on Good Cell Culture Practice*, Vol. 33, pp.261–287, a report of the second ECVAM task force on good cell culture Practice, Alternatives to Laboratory Animals – ATLA.

Dabrazhynetskaya, A.A., Volokhov, D.V.D. and David, S.W.S. et al. (2011) 'Preparation of reference strains for validation and comparison of mycoplasma testing methods', *Journal of Applied Microbiology*, Vol. 111, No. 4, pp.904–914.

Dietrich, D.R. (2010) 'Courage for simplification and imperfection in the 21st century assessment of 'Endocrine disruption', *ALTEX*, Vol. 27, No. 4, pp.264–278.

DiMasi, J.A. and Grabowski, H.G. (2007) 'The cost of biopharmaceutical R&D: is biotech different?', *Manage. Decis. Econ.*, Vol. 28, Nos. 4–5, pp.469–479.

Dirks, W.G., Macleod, R.A.F. and Nakamura, Y. et al. (2010) 'Cell line cross-contamination initiative: an interactive reference database of STR profiles covering common cancer cell lines', *International Journal of Cancer*, Vol. 126, No. 1, pp.303–304.

Drexler, H.G. and Uphoff, C.C. (2002) 'Mycoplasma contamination of cell cultures: incidence, sources, effects, detection, elimination, prevention', *Cytotechnology*, Vol. 39, No. 2, pp.75–90.

EFSA GMO Panel Working Group on Animal Feeding Trials (2008) 'Safety and nutritional assessment of GM plants and derived food and feed: the role of animal feeding trials', *Food Chem. Toxicol.*, Vol. 46, No. Suppl. 1, pp.S2–S70.

Fang, F.C., Steen, R.G. and Casadevall, A. (2012) 'Misconduct accounts for the majority of retracted scientific publications', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109, No. 42, pp.17028–17033.

Festing, M. (2003) 'Principles: the need for better experimental design', *Trends in Pharmacological Sciences*, Vol. 27, No. 7, pp.341–345.

Gilbert, J., Henske, P. and Singh, A. (2003) 'Rebuilding big pharma's business model, in vivo, the business & medicine report', *Windhover Information*, Vol. 21, No. 10 [online] https://www.ucl.ac.uk/wibr/teaching/Docs/rebuilding_big_pharma.pdf.

Gocht, T., Berggren, E., Ahr, H.J. et al. (2015) 'The SEURAT-1 approach towards animal free human safety assessment', *ALTEX*, Vol. 32, No. 1, pp.9–24.

Gold, L.S., Manley, N.B. and Slone, T.H. et al. (2005) 'Supplement to the carcinogenic potency database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997–1998', *Toxicol. Sci.*, Vol. 85, No. 2, pp.747–808.

Hackam, D.G. (2007) 'Translating animal research into clinical benefit', *BMJ: British Medical Journal*, Vol. 334, No. 7586, pp.163–164.

Hackam, D.G. and Redelmeier, D.A. (2006) 'Translation of research evidence from animals to humans', *JAMA: the Journal of the American Medical Association*, Vol. 296, No. 14, pp.1731–1732.

Hartung, T. (2007a) 'Food for thought ... on cell culture', *ALTEX*, Vol. 24, No. 3, pp.143–152.

Hartung, T. (2007b) 'Food for thought … on validation', *ALTEX*, Vol. 24, No. 2, pp.67–72.

Hartung, T. (2008a) 'Food for thought ... on alternative methods for cosmetics safety testing', *ALTEX*, Vol. 25, No. 3, pp.147–162.

Hartung, T. (2008b) 'Food for thought ... on animal tests', *ALTEX*, Vol. 25, No. 3, pp.3–16.

Hartung, T. (2008c) 'Toward a new toxicology – evolution or revolution?', *ATLA – Alternatives to Laboratory Animals*, Vol. 36, No. 6, pp.635–639.

Hartung, T. (2009a) 'Food for thought ... on evidence-based toxicology', *ALTEX*, Vol. 26, No. 2, pp.75–82.

Hartung, T. (2009b) 'Toxicology for the twenty-first century', *Nature*, Vol. 460, No. 7252, pp.208–212.

Hartung, T. (2009c) 'Per aspirin ad astra...', *Alternatives to Laboratory Animals: ATLA*, Vol. 37, No. Suppl. 2, pp.45–47.

Hartung, T. (2010a) 'Comparative analysis of the revised Directive 2010/63/EU for the protection of laboratory animals with its predecessor 86/609/EEC – a t4 report', *ALTEX*, Vol. 27, No. 4, pp.285–303.

Hartung, T. (2010b) 'Evidence-based toxicology – the toolbox of validation for the 21st century?', *ALTEX*, Vol. 27, No. 4, pp.253–263.

Hartung, T. (2010c) 'Food for thought ... on alternative methods for nanoparticle safety testing', *ALTEX*, Vol. 27, No. 2, pp.87–95.

Hartung, T. (2010d) 'Food for thought ... on alternative methods for chemical safety testing', *ALTEX*, Vol. 27, No. 1, pp.3–14.

Hartung, T. (2010e) 'Lessons learned from alternative methods and their validation for a new toxicology in the 21st century', *J. Toxicol. Env. Health*, Vol. 13, Nos. 2–4, pp.277–290.

Hartung, T. (2013) 'Look back in anger – what clinical studies tell us about preclinical work', *ALTEX*, Vol. 30, No. 3, pp.275–291.

Hartung, T. and Corsini, E. (2013) 'Immunotoxicology: challenges in the 21st century and in vitro opportunities', *ALTEX*, Vol. 30, No. 4, pp.411–426.

Hartung, T. and Daston, G. (2009) 'Are in vitro tests suitable for regulatory use?', *Tox. Sci.*, Vol. 111, No. 2, pp.233–237.

Hartung, T. and Hoffmann, S. (2009) 'Food for thought ... on in silico methods in toxicology', *ALTEX*, Vol. 26, No. 3, pp.155–166.

Hartung, T. and Koeter, H. (2008) 'Food for thought … on alternative methods for food safety testing', *ALTEX*, Vol. 25, No. 4, pp.259–264.

Hartung, T. and McBride, M. (2011) 'Food for thought … on mapping the human toxome', *ALTEX*, Vol. 28, No. 2, pp.83–93.

Hartung, T. and Rovida, C. (2009) 'Chemical regulators have overreached', *Nature*, Vol. 460, No. 7259, pp.1080–1081.

Hartung, T. and Sabbioni, E. (2011) 'Alternative in vitro assays in nanomaterial toxicology', *Nanomedicine and Nanobiotechnology*, Vol. 3, No. 6, pp.545–573, DOI: 10.1002/wnan.153.

Hartung, T. and Zurlo, J. (2012) 'Food for thought ... Alternative approaches for medical countermeasures to biological and chemical terrorism and warfare', *ALTEX*, Vol. 29, No. 3, pp.251–260.

Hartung, T., Bremer, S. and Casati, S. et al. (2004) 'A modular approach to the ECVAM principles on test validity', *ATLA – Alternatives to Laboratory Animals*, Vol. 32, No. 5, pp.467–472.

Hartung, T., van Vliet, E. and Jaworska, J. et al. (2012) 'Food for thought ... systems toxicology', *ALTEX*, Vol. 29, No. 2, pp.119–128.

Hay, R.J., Macy, M.L. and Chen, T.R. (1989) 'Mycoplasma infection of cultured cells', *Nature*, Vol. 339, No. 6224, pp.487–488.

Hoffmann, S. and Hartung, T. (2006) 'Toward an evidence-based toxicology', *Human and Experimental Toxicology*, Vol. 25, Nos. 2–3, pp.497–513.

Hoffmann, S., Stephens, M. and Hartung, T. (2014) 'Evidence-based toxicology', in Wexler, P. (Ed.): *Encyclopedia of Toxicology*, 3rd ed., Vol. 2, pp.565–567, Elsevier Inc., Academic Press.

Hooijmans, C.R., Leenaars, M. and Ritskes-Hoitinga, M. (2010) 'A gold standard publication checklist to improve the quality of animal studies, to fully integrate the three Rs, and to make systematic reviews more feasible', *ATLA – Alternatives to Laboratory Animals*, Vol. 38, No. 2, pp.167–182.

Hughes, P., Marshall, D., Reid, Y., Parkes, H. and Gelber, C. (2007) 'The costs of using unauthenticated, over-passaged cell lines: how much more data do we need?', *BioTechniques*, Vol. 43, No. 5, pp.575–584.

Ioannidis, J.P.A. (2005a) 'Contradicted and initially stronger effects in highly cited clinical research', *JAMA: The Journal of the American Medical Association*, Vol. 294, No. 2, pp.218–226.

Ioannidis, J.P.A. (2005b) 'Why most published research findings are false', *PLoS Medicine*, Vol. 2, No. 8, pp.e124–e124.

Janssens, A.C.J.W. and van Duijn, C.M. (2008) 'Genome-based prediction of common diseases: advances and prospects', *Human Molecular Genetics*, Vol. 17, No. R2, pp.R166–R173.

Juberg, D.R., Borghoff, S.J. and Becker, R.A. et al. (2013) 't4 workshop report: lessons learned, challenges, and opportunities: the US endocrine disruptor screening program', *ALTEX*, Vol. 31, No. 6, pp.63–78.

Kilkenny, C.C., Browne, W.J., Cuthill, I.C., Emerson, M. and Altman, D.G. (2010) 'Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research', *PLoS Biology*, Vol. 8, No. 6, p.e1000412.

Kipen, H.M. and Laskin, D.L. (2005) 'Smaller is not always better: nanotechnology yields nanotoxicology', *Am. J. Physiol. Lung. Cell. Mol. Physiol.*, Vol. 289, No. 5, pp.L696–L697.

Knight, A. (2007) 'Systematic reviews of animal experiments demonstrate poor human clinical and toxicological utility', *ATLA – Alternatives to Laboratory Animals*, Vol. 35, No. 5, pp.641–659.

Knight, J. and Rovida, C (2014) 'Safety evaluations under the proposed US Safe Cosmetics and Personal Care Products Act of 2013: animal use and cost estimates', *ALTEX*, Vol. 31, No. 2, pp.177–208.

Landry, J.J.M., Pyl, P.T. and Rausch, T. et al. (2013) 'The genomic and transcriptomic landscape of a HeLa cell line', *G3 Genes-Genomes-Genetics Early Online*, published on 11 March, DOI: 10.1534/g3.113.005777.

Lang, T. (2004) 'Twenty statistical errors even you can find in biomedical research articles', *Croatian Medical Journal*, Vol. 45, No. 4, pp.361–370.

Langdon, S.P. (2003) 'Cell culture contamination: an overview', in Langdon, S.P. (Ed.): *Methods in Molecular Medicine, Cancer Cell Culture: Methods and Protocols Cancer Cell Culture*, Vol. 88, pp.309–318, Humana Press, Totowa, New Jersey.

Lawrence, B., Bashiri, H. and Dehghani, H. (2010) 'Cross comparison of rapid mycoplasma detection platforms', *Biologicals*, Vol. 38, No. 2, pp.6–6.

Leist, M. and Hartung, T. (2013) 'Reprint: inflammatory findings on species extrapolations: humans are definitely no 70-kg mice', *ALTEX*, Vol. 30, No. 2, pp.227–230.

Leist, M., Efremova, L. and Karreman, C. (2010) 'Food for thought ... considerations and guidelines for basic test method descriptions in toxicology', *ALTEX*, Vol. 27, No. 4, pp.309–317.

Leist, M., Hasiwa, M., Daneshian, M. and Hartung, T. (2012) 'Validation and quality control of replacement alternatives – current status and future challenges', *Toxicological Research*, Vol. 1, pp.8–22, DOI: 10.1039/C2TX20011B.

Leist, M., Hasiwa, N., Rovida, C. et al. (2014) 'Consensus report on the future of animal-free systemic toxicity testing', *ALTEX*, Vol. 31, No. 3, pp.341–356.

Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H. and Hartung, T. (2016) 'Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008–2014', *ALTEX*, Vol. 33, No. 2, pp.95–109 [online] http://doi.org/10.14573/altex.1510052.

Lumbreras, B., Parker, L.A., Porta, M., Pollan, M., Ioannidis, J.P.A. and Hernandez-Aguado, I. (2009) 'Overinterpretation of clinical applicability in molecular diagnostic research', *Clinical Chemistry*, Vol. 55, No. 4, pp.786–794.

Lund, R.J., Närvä, E. and Lahesmaa, R. (2012) 'Genetic and epigenetic stability of human pluripotent stem cells', *Nature Reviews Genetics*, Vol. 13, No. 10, pp.732–744.

MacCallum, C.J. (2010) 'Reporting animal studies: good science and a duty of care', *PLoS Biol.*, Vol. 8, p.e1000413, DOI: 10.1371/journal.pbio.1000413.

Macleod, M. and van der Worp, H.B. (2010) 'Animal models of neurological disease: are there any babies in the bathwater?', *Practical Neurolology*, Vol. 10, No. 6, pp.312–314.

Macleod, R.A.F., Dirks, W.G., Matsuo, M., Kaufmann, Y., Milch, H. and Drexler, H.G. (1999) 'Widespread intraspecies cross-contamination of human tumor cell lines arising at source', *International Journal of Cancer*, Vol. 83, No. 4, pp.555–563.

Maertens, A., Anastas, N., Spencer, P.J., Stephens, M., Goldberg, A. and Hartung, T. (2014) 'Green toxicology', *ALTEX*, Vol. 31, No. 3, pp.243–249.

Mignini, L.E. and Khan, K.S. (2006) 'Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research', *BMC Medical Research Methodology*, Vol. 6, p.10, DOI: 10.1186/1471-2288-6-10.

Mitalipova, M.M., Rao, R.R. and Hoyer, D.M. et al. (2005) 'Preserving the genetic integrity of human embryonic stem cells', *Nature Biotechnology*, Vol. 23, No. 1, pp.19–20.

Mundae, M.K. and Östor, A.J.K. (2010) 'The long road of biopharmaceutical drug development: from inception to marketing', *Q.J. Med.*, Vol. 103, No. 1, pp.3–7.

Neltner, T.G., Alger, H.M. and Leonard, J.E. et al. (2013) 'Data gaps in toxicity testing of chemicals allowed in food in the United States', *Reproductive Toxicology*, Vol. 42, pp.85–94.

Nikfarjam, L. and Farzaneh, P. (2012) 'Prevention and detection of Mycoplasma contamination in cell culture', *Cell Journal (Yakhteh)*, Vol. 13, No. 4, pp.203–212.

NRC – National Research Council, Committee on Animal Models for Assessing Countermeasures to Bioterrorism Agents (2011) *Animal Models for Assessing Countermeasures to Bioterrorism Agents*, pp.1–153, the National Academies Press, Washington, DC, USA [online] http://dels.nationalacademies.org/Report/Animal-Models-Assessing-Countermeasures/13233 (accessed 20 April 2016).

OECD (2005) *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*, OECD Series on testing and assessment No. 34, pp.1–96.

Olson, H., Betton, G. and Robinson, D. et al. (2000) 'Concordance of the toxicity of pharmaceuticals in humans and in animals', *Regulatory Toxicology and Pharmacology*, Vol. 32, No. 1, pp.56–67.

Pajares, F. (n.d.) *Outline and Study Guide to the Structure of Scientific Revolutions by Thomas S. Kuhn*, Emory University, Atlanta, GA, USA [online] http://www.des.emory.edu/mfp/Kuhn.html (accessed 15 October 2008).

Patlewicz, G., Ball, N., Becker, R.A., Blackburn, K., Booth, E., Cronin, M., Kroese, D., Steup, D., van Ravenzwaay, B. and Hartung, T. (2014) 'Read-across approaches – misconceptions, promises and challenges ahead', *ALTEX*, Vol. 31, No. 4, pp.387–396.

Perel, P., Roberts, I. and Sena, E. et al. (2007) 'Comparison of treatment effects between animal experiments and clinical trials: systematic review', *BMJ (Clinical Research Ed.)*, Vol. 334, No. 7586, pp.197–197, doi:10.1136/bmj.39048.407928.BE.

Pottenger, L.H., Bus, J.S. and Gollapudi, B.B. (2007) 'Genetic toxicity assessment: employing the best science for human safety evaluation part VI: when salt and sugar and vegetables are positive, how can genotoxicity data serve to inform risk assessment?', *Toxicological Sciences*, Vol. 98, pp.327–331.

Pound, P., Ebrahim, S., Sandercock, P., Bracken, M.B. and Roberts, I. (2004) 'Reviewing animal trials systematically (RATS) group: where is the evidence that animal research benefits humans?', *BMJ (Clinical Research Ed.)*, Vol. 328, pp.514–517.

Prinz, F., Schlange, T. and Asadullah, K. (2011) 'Believe it or not: how much can we rely on published data on potential drug targets?', *Nature Reviews Drug Discovery*, Vol. 10, No. 9, pp.712–712.

Reisinger, K., Hoffmann, S., Alepée, N. et al. (2015) 'Systematic evaluation of non-animal test methods for skin sensitisation safety assessment', *Toxicol. In Vitro*, Vol. 29, No. 1, pp.259–270.

Roberts, I., Kwan, I., Evans, P. and Haig, S. (2002) 'Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation', *BMJ: British Medical Journal*, Vol. 324, pp.474–476.

Rojas, A., Gonzalez, I. and Figueroa, H. (2008) 'Cell line cross-contamination in biomedical research: a call to prevent unawareness', *Acta Pharmacologica Sinica*, Vol. 29, No. 7, pp.877–880.

Rottem, S. and Barile, M.F. (1993) 'Beware of mycoplasmas', *Trends in Biotechnology*, Vol. 11, No. 4, pp.143–151.

Rovida, C. and Hartung, T. (2009) 'Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals – a report by the transatlantic think tank for toxicology (t4)', *ALTEX*, Vol. 26, No. 4, pp.187–208.

Rovida, C., Asakura, C., Daneshian, M. et al. (2015) 'Toxicity testing in the 21st century beyond environmental chemicals', *ALTEX*, Vol. 32, No. 1, pp.25–40.

Rovida, C., Longo, F.F. and Rabbit, R.R. (2011) 'How are reproductive toxicity and developmental toxicity addressed in REACH dossiers?', *ALTEX*, Vol. 28, No. 3, pp.273–294.

Russell, W.M.S. and Burch, R.L. (1959) *The Principles of Humane Experimental Technique*, Methuen, London, UK [online] http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc (accessed 20 April 2016).

Scannell, J.W., Blanckley, A., Boldon, H. et al. (2012) 'Diagnosing the decline in pharmaceutical R&D efficiency', *Nature Rev. Drug Discovery*, Vol. 11, No. 3, pp.191–200.

Seaton, A. and Donaldson, K. (2005) 'Nanoscience, nanotoxicology, and the need to think small', *The Lancet*, Vol. 365, No. 9463, pp.923–924.

Seok, J., Warren, H.S. and Cuenca, A.G. et al. (2013) 'Genomic responses in mouse models poorly mimic human inflammatory diseases', *Proceedings of the National Academy of Science USA*, Vol. 110, No. 4, pp.3507–3512.

Skloot, R. (2010) *The Immortal Life of Henrietta Lacks*, Reprint edition, 402pp, Crown, New York.

Smirnova, L., Hogberg, H.T., Leist, M. and Hartung, T. (2014) 'Developmental neurotoxicity – challenges in the 21st century and in vitro opportunities', *ALTEX*, Vol. 31, No. 2, pp.129–156.

Stacey, G.N. (2000) 'Cell contamination leads to inaccurate data: we must take action now', *Nature*, Vol. 203, No. 6768, p.356.

Steinemann, D., Göhring, G. and Schlegelberger, B. (2013) 'Genetic instability of modified stem cells – a first step towards malignant transformation?', *American Journal of Stem Cells*, Vol. 2, No. 1, pp.39–51.

Stephens, M.L., Andersen, M., Becker, R.A. et al. (2013) 'Evidence-based toxicology for the 21st century: opportunities and challenges', *ALTEX*, Vol. 30, No. 1, pp.74–104.

Tamimi, N.A.M. and Ellis, P. (2009) 'Drug development: from concept to marketing!', *Nephron. Clin. Pract.*, Vol. 13, No. 3, pp.c125–c131.

van der Worp, H.B. and Macleod, M.R. (2011) 'Preclinical studies of human disease: time to take methodological quality seriously', *Journal of Molecular and Cellular Cardiology*, Vol. 51, No. 4, pp.449–450.

van der Worp, H.B., Howells, D.W. and Sena, E.S. et al. (2010) 'Can animal models of disease reliably inform human studies?', *PLoS Med.*, Vol. 7, p.e1000245.

Young, L., Sung, J., Stacey, G. and Masters, J.R. (2010) 'Detection of mycoplasma in cell cultures', *Nature Protocols*, Vol. 5, No. 5, pp.929–934.

Young, N.S., Ioannidis, J.P.A. and Al-Ubaydli, O. (2008) 'Why current publication practices may distort science', *PLoS Medicine*, Vol. 5, No. 10, p.e201.

Zurlo, J. (2011) 'Evidence-based toxicology collaboration kick-off meeting', *ALTEX*, Vol. 28, No. 2, p.152.

Notes

[1] http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/.

[2] http://en.wikipedia.org/wiki/List_of_withdrawn_drugs.

[3] http://www.imi.europa.eu/content/mip-dili.

[4] https://eurl-ecvam.jrc.ec.europa.eu/about-ecvam/archive-publications/publication/ESAC27_statement_ACUTE.pdf.

[5] http://www.nature.com/news/most-popular-human-cell-in-science-gets-sequenced-1.12609.

[6] http://www.hpacultures.org.uk/services/celllineidentityverification/misidentifiedcelllines.jsp.

[7] http://ebtox.com/.