Perspectives on *In Vitro* to *In Vivo* Extrapolations

Thomas Hartung

Abstract

Quantitative *in vitro* to *in vivo* extrapolation (QIVIVE) is broadly considered a prerequisite bridge from *in vitro* findings to a dose paradigm. Quality and relevance of cell systems are the first prerequisite for QIVIVE. Information-rich and mechanistic endpoints (biomarkers) improve extrapolations, but a sophisticated endpoint does not make a bad cell model a good one. The next need is reverse toxicokinetics (TK), which estimates the dose necessary to reach a tissue concentration that is active *in vitro*. The Johns Hopkins Center for Alternatives to Animal Testing (CAAT) has created a roadmap for animal-free systemic toxicity testing, in which the needs and opportunities for TK are elaborated, in the context of different systemic toxicities. The report was discussed at two stakeholder forums in Brussels in 2012 and in Washington in 2013; the key recommendations are summarized herein. Contrary to common belief and the Paracelsus paradigm of everything is toxic, the majority of industrial chemicals do not exhibit toxicity. Strengthening the credibility of negative results of alternative approaches for hazard identification, therefore, avoids the need for QIVIVE. Here, especially the combination of methods in integrated testing strategies is most promising. Two further but very different approaches aim to overcome the problem of modeling *in vivo* complexity: The *human-on-a-chip* movement aims to reproduce large parts of living organism's complexity via microphysiological systems, that is, organ equivalents combined by microfluidics. At the same time, the Toxicity Testing in the 21st Century (Tox-21c) movement aims for mechanistic approaches (adverse outcome pathways as promoted by Organisation for Economic Co-operation and Development (OECD) or pathways of toxicity in the Human Toxome Project) for high-throughput screening, biological phenotyping, and ultimately a systems toxicology approach through integration with computer modeling. These 21st century approaches also require 21st century validation, for example, by evidence-based toxicology. Ultimately, QIVIVE is a prerequisite for extrapolating Tox-21c such approaches to human risk assessment.

Keywords: alternative methods, extrapolation, new approaches, validation

> Where is the wisdom
> we have lost in knowledge?
> And where is the knowledge
> we have lost in information?
> T.S. Eliot

Introduction

MERRIAM-WEBSTER DEFINES extrapolation as ''to project, extend, or expand (known data or experience) into an area not known or experienced so as to arrive at a usually conjectural knowledge of the unknown area.'' In toxicology, extrapolation is the logical consequence of using models and not the systems themselves to study any phenomenon.[1] As toxicologists, we want to protect real people regardless of gender, life stage and life/health history, genetic background, and lifestyle from exposures as well as complex mixed exposures. An interim step is the assumption of a human hazard, perhaps a healthy 70 kg (male?) young adult exposed to a single substance in a single dose or steady exposure (Fig. 1). Typically, we are not very explicit with this construct, but the assumption that some general truth can be deduced is implicit; the individual and circumstantial conditions create the deviations and individual susceptibilities and, in the end, the uncertainties. This is best described as distributions and probabilities. The main sources for approximating ''Human Hazard'' are epidemiological/clinical, animal studies, human-on-a-chip approaches, and *in vitro* models with different degrees of integration (including toxicokinetic aspects).

This article summarizes some personal lessons learned by the author on how quantitative *in vitro* to *in vivo* extrapolation (QIVIVE)[2] from good cell models (especially organo-typic cultures and human-on-a-chip models) is key to allow their use for implementing a mechanistic toxicology

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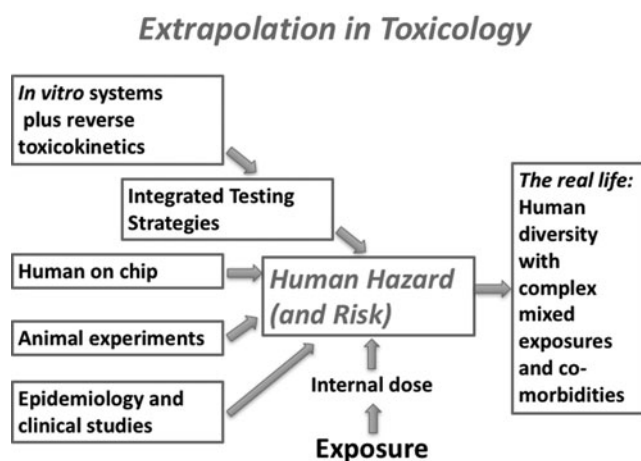University of Konstanz, Konstanz, Germany.

Extrapolation in Toxicology

FIG. 1. Different information sources feed the extrapolation to human hazard, a construct from which to extrapolate to real-life situations.

as called for by the Toxicology for the 21st Century Movement.[3,4] A view on the quality needs of the different information sources forms the starting point of these attempts. Mechanistic understanding organized by adverse outcome pathways[5] (AOP) is key for the integration and extrapolation of *in vitro* information, typically from different information sources as Integrated Testing Strategies[6,7] (ITS or Integrated Approaches to Testing and Assessment [IATA], as now called by Organisation for Economic Co-operation and Development [OECD][8]). However, for a systems toxicology approach[9–11] based on modeling the organism's behavior under stress of a toxicant, more refined molecular and quantitative pathways of toxicity (PoT) are required. However, we also need other ways of further integrating the existing information and a case is made for systematic reviews and evidence-based toxicology (EBT).[12]

The Experimental Models and Studies Forming the Basis for Extrapolation to Human Hazard

Epidemiology and clinical studies

Observational and experimental studies on humans, particularly in toxicology, are limited by their costs and ethical considerations. Frequently, instead of the obviously desirable controlled experiment, involuntary exposures (e.g., at the workplace) are studied, though only retrospectively. Interestingly, in contrast to the other approaches discussed herein (which suffer from a lack of variability of study subjects, representing real life), epidemiology and clinical studies suffer from exactly this variability of humans and have to distill the underlying hazard and risk. The question here is whether findings can be generalized, that is, extrapolated to other populations and exposure scenarios. Where available, they constitute an important part of the construct of "Human Hazard," but they will not be covered in this article except as possible points of reference for validation.

Animal studies

Over the last 100 years, *in vivo* studies have been the primary source of information in toxicology. The data have shaped our view on toxic substances and we could easily make this world a safer place—for rats and mice. There are many reasons to discard this approach, including ethical and economical, but the lack of human relevance is the most important.[13,14] Table 1 (reproduced from an article[15]) summarizes variations in animal studies from human clinical trials, illustrating the associated uncertainty and the need and challenge to extrapolate from animal studies to humans. This is especially relevant when animal studies are used as points of reference for validation of *in vitro* systems,[16] which by definition then cannot overcome the shortcomings of the animal test.[17–19] Ironically, where the *in vitro* system is better than the animal, it will be held against it as an inaccurate prediction of the *in vivo* result. Some recent work has shown based on large data sets, the reproducibility issues of animal studies for eye irritation[20,21] and skin sensitization[22–24] as pertinent examples.

Animal studies require intraspecies extrapolation, first from the model to humans and also among different animal species themselves—in veterinary drug development, for example, which might require intraspecies extrapolations, a Yorkshire Terrier is quite different from a Beagle. Where there are few pharmacological target differences, that is, when substances interact with biology (for small molecules—biologicals are very different), some differences in metabolism and a large number of variations in kinetics should be expected. Thus, toxicokinetics (TK) plays a key role. Defense reactions (resilience)[25] might actually vary more than the initial vulnerability of the different species.

Traditional *in vitro systems*

Any *in vitro* system—even human-on-a-chip approaches based on cutting edge science and bioengineering—is a tremendous simplification of the human body. A well-known textbook, *Molecular Biology of the Cell* by Alberts, Johnson, Lewis, Morgan, Raff, Roberts, and Walter, lists more than 200 different cell types[26] and many of these come in variants and subtypes. It is impossible to represent them all under reasonable physiological culture conditions or enable their interactions in any model system. Thus, there will always be extrapolation involved if we need to predict how the entire organism would have reacted. It is likely that many cell types are not critically involved in the PoT or AOP and we can expect meaningful results from a more simplified approach. In some cases, there will be trigger points (molecular initiating events) or crucial steps in a sequence (key events), which sufficiently reflect the mechanism and allow very simple systems to test them. This is extending the *International Programme on Chemical Safety* Mode of Action framework[27,28] in animals to address human relevance to cellular models. Both *in silico* and *in vitro* models are most promising to predict adverse outcomes, when they reflect these molecular initiating and key events. There will always be tradeoff between the complexity and completeness of the model and the quality of extrapolation.[29] Each element we omit might represent the entry port of a yet unknown modulating effect. We should therefore follow Albert Einstein's adage: "Everything should be made as simple as possible, but not simpler." How simple the input can be will depend very much on the quality of the input parameters and the extrapolation.

We should be clear that typical cell models have many limitations (as discussed in Refs.[15,30,31]). A few prominent ones include the following:

TABLE 1. DIFFERENCES BETWEEN AND METHODOLOGICAL PROBLEMS OF ANIMAL AND HUMAN STUDIES
CRITICAL TO PREDICTION OF SUBSTANCE EFFECTS

Subjects
  Small groups of (often inbred, homogenous genetical background) animals vs. large groups of individuals with
    heterogeneous genetical background
  Young adult animals vs. all ages in human trials
  Animals typically only one gender
  Disparate animal species and strains, with a variety of metabolic pathways and drug metabolites, leading to variation in
    efficacy and toxicity
Disease models
  Artificial diseases, i.e., different models for inducing illness in healthy animals or injury with varying similarity to the
    human condition of sick people
  Acute animal models for chronic phenomena
  Monofactorial disease models vs. multifactorial ones in humans
  Especially in knock-out mouse models the adaptive responses in animals are underestimated compensating for the knock-out
Doses
  Variations in drug dosing schedules (therapeutic to toxic) and regimen (usually once daily) that are of uncertain relevance
    to the human condition (therapeutic optimum)
  Pharmaco- and toxicokinetics of substances differ between animals and humans
Circumstances
  Uniform, optimal housing and nutrition vs. variable human situations
  Animals are stressed
  Never concomitant therapy vs. frequent ones in humans
Diagnostic procedures
  No vs. intense verbal contact
  Limited vs. extensive physical exam in humans
  Limited standardized vs. individualized clinical laboratory examination in humans
  Predetermined timing vs. individualized in humans
  Extensive histopathology vs. exceptional one in humans
  Length of follow up before determination of disease outcome varies and may not correspond to disease latency in humans
  Especially in toxicological studies the prevalence of health effects are rarely considered when interpreting data
Study design
  Variability in the way animals are selected for study, methods of randomisation, choice of comparison therapy (none,
    placebo, vehicle), and reporting of loss to follow up
  Small experimental groups with inadequate power, simplistic statistical analysis that does not account for potential
    confounding, and failure to follow intention to treat principles
  Nuances in laboratory technique that may influence results may be neither recognised nor reported—e.g., methods for
    blinding investigators
  Selection of a variety of outcome measures, which may be disease surrogates or precursors and which are of uncertain
    relevance to the human clinical condition
  Traditional designs especially of guideline studies offering standardization but prohibiting progress

Combined from Olson et al. (2000),[32] Pound et al. (2004),[33] and Hartung (2008),[13] reproduced from Hartung (2013)[15] with permission.

- Mycoplasma: unacceptably, many cultures are not controlled for infections; the availability and higher control of cell culture reagents have improved the situation, but many human respiratory infection-derived mycoplasmas are found in cultures.
- Dedifferentiation favored by growth conditions and cell selection—we want to expand our cells and choose the precise conditions that contradict differentiation.
- Many cell functions are stimulated during culture—cell functions and cell mass are only maintained when needed. We pamper our (primary) cells and are surprised that they do not maintain functions.
- Lack of oxygen: normal dense cell cultures consume the oxygen dissolved in the medium within a few hours and need to use anaerobic metabolism.
- Lack of metabolism and defense: the former is often discussed in toxicology, where it is sometimes the metabolite that triggers the insult.[34] The problem seems to

be solved mostly for drugs, but we do not use these rather resource-intensive methods for environmental chemicals. Before this, it might be better to have no metabolism than the wrong metabolism, as it is usually protective and might hide a problem if the animals metabolize differently. However, cells also lose defense capabilities in culture, rendering them more sensitive to toxic insult that might exaggerate problems.
- Unknown fate of test compounds in culture: test agents have kinetics in culture (e.g., solubility, distribution, binding, and chemical reactions), which we typically neglect.
- Tumor origin of many cells: dramatic genetic instability is still frequently not taken into consideration, contributing to irreproducible results.[35]
- Cell identity: it is difficult to believe, but work is still carried out quite frequently (up to 25% of studies) with cell lines that are not what they are supposed to be.

For this reason, the importance of quality assurance measures such as good cell culture practice (GCCP[36,37]) cannot be stressed enough. These efforts are continuing, for example, with a GCCP for induced pluripotent stem cells (iPSC) workshop in spring 2015[38] and Center for Alternatives to Animal Testing (CAAT) work on reporting standards for *in vitro* work. Notably, there are other activities in process such as the *Good In vitro Method Practice* (GIVIMP) by ECVAM and the OECD, which have been recently published.[39] Furthermore, some of the more organotypic culture techniques now emerging may be of assistance,[40,41] providing more relevant models with adequate cell densities/cell contacts in 3D, cocultures of relevant cells, and homeostatic environments by perfusion, among others. Simply stated, the more relevant the model, the more likely extrapolation will be correct.

How to extrapolate now to human complexity? Two current approaches (Fig. 2) to move beyond simple *in vitro* tests as full substitutes for animal tests are as follows:

1. Increased resolution of (molecular) pathways aiming for a mechanistic toxicology and ultimately for systems biology/toxicology/pharmacology and virtual models.
2. Reproducing complexity in the experimental system, which reaches from organotypic 3D cultures to human-on-chip systems.

Pathway-based toxicological models

Under the label of Toxicity Testing for the 21st Century[42–45] and more recently AOP,[46] numerous activities have furthered testing based on toxicity mechanisms (see also the PoT as Basis of Extrapolation section). This is, however, easier said than done[47]: What is a toxicity pathway? How do we annotate them? How complete is our picture of these pathways? What is causal, what is not? What are the thresholds of adversity in such pathway perturbations? How conserved are these pathways between cells, species, and hazards? Can we produce consensus and a repository of pathways?

Two basic assumptions drive this approach: (a) there are mechanisms that are distinct and conserved and (b) there are not too many of them. Both assumptions do not necessarily hold true. We are dealing with extremely complex, networked systems, which contrast with our mostly linear views of the sequences of events and independence of key events. The presence or abundance of a single component of the pathway may result in differences of pathway perturbation. Can we then speak of a defined pathway at all?

Microphysiological systems

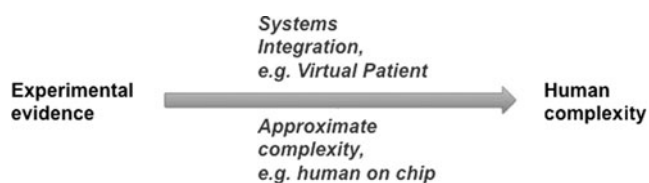Microphysiological systems (MPS) is a term increasingly used for more physiological cultures[48] that make (to different degrees) use of 3D culture, stem cell-derived tissues, scaffolds, extracellular matrix, cell and organoid coculture, perfusion, oxygen supply, physical stretch, and so on. This could overcome many test shortcomings, especially when using stem cells.[41] The increasing use of the phrase "human-on-chip" reflects the ambitions to make such systems as close to the entire organism as possible. The recent stimulus in research bringing together cell cultures and bioengineering was prompted by the desire of the U.S. Department of Defense to develop medical countermeasures for biological and chemical warfare and terrorism.[49] The lack of patients and adequate animal models—as highlighted by the evaluation of a National Academy of Sciences panel the author contributed to[50]—led to a $200 million program by National Institutes of Health/Food and Drug Administration/Defense Advanced Research Projects Agency, Defense Threat Reduction Agency[49] and, most recently, the Environmental Protection Agency in the United States.

Our own work in this context led to the development of a human mini-brain from iPSC[50–53] for developmental neurotoxicity testing.[54] This is only one example of rapidly developing 3D culture systems[41] that often make use of human stem cells combined with bioengineering to form MPS.[48,55] One goal of Tox-21c is the deduction of PoT. The complex mixture of cell types in the mini-brain model and the ongoing maturation of the system can be a disadvantage here. We therefore developed in parallel a model of 3D shaker cultures of dopaminergic neurons derived from Lund Human Mesencephalic (LUHMES) cells.[56] This model has been used to identify PoT of toxicants such as MPP$^+$ and rotenone.

MPS that combine different iPSC-derived cell types in a specific 3D configuration into mini-organoids to generate a human-on-a-chip could enable studies of complex cellular networks and disease models for drug development, toxicology, and medicine. Many perfusion platforms have been developed and the human mini-brain model is currently combined with some of these platforms. These examples show that Tox-21c also needs a cell culture for the 21st century, that is, one that is more organotypic and quality assured.

Toxicokinetics

TK is the necessary complement to ALL *in vitro* approaches.[57] TK in animal studies is not a stand-alone test requirement in most regulatory schemes, but represents the necessary complement to *in vitro* approaches as it indicates whether meaningful concentrations are tested and how they relate to doses applied (the paradigm of our current risk assessment). The most advanced approach is physiology-based pharmacokinetic (PBPK) modeling.[58] What QIVIVE needs is a type of retro-PBPK. While most PBPK is done to model tissue concentrations over time after administration *in vivo*, retro-PBPK, as approached over the last few years, calculates the doses, which would result in the local concentrations effective *in vitro*. The distinction of "external" and "internal exposure" is useful to distinguish what the organism versus its tissues is exposed to, that is, the result of TK.

In a series of meetings, we developed a roadmap for animal-free systemic toxicity testing.[59,60] This consensus was quite optimistic about the potential to avoid animal testing for TK in its evaluations and recommendations:

FIG. 2. The two paths for approximating human complexity from experimental evidence.

- *In silico* approaches have to be optimized: PBPK modeling platforms need to be more user friendly and open source; physiological parameters (dermal, inhalation exposure, etc.) need to be curated.
- Data collection for quantitative structure/activity relationship modeling and simulation, including metabolism, distribution, and protein binding that allow for simulation based on physicochemical properties, needs to be created.
- Data collections are further needed to support QIVIVE, which should incorporate *in vitro* data (barrier models, e.g., placenta, mammary, testis, intestine, and brain, as well as specific transporters).
- Problems are lacking models for bioavailability and urinary excretion, nonhepatic metabolism, blood/brain barriers, and gastrointestinal metabolism.
- Free concentrations/target concentrations *in vitro* need to be used for extrapolation.
- However, these challenges seem to be—with reasonable investment—achievable.

We have most recently summarized the challenges and opportunities of putting this into practice.[2]

Biomarkers—The Meaningful Endpoints to Measure

Extrapolation depends as much on the model system it does on what we measure to characterize the response. The term biomarker is defined as follows[61]: "Indicator signaling an event or condition in a biological system or sample and giving a measure of exposure, effect, or susceptibility." We addressed the biomarkers of *in vitro* systems in an earlier workshop.[62] The key conclusion was that biomarkers need to reflect the mode of action/mechanism that distinguish them from the many other things happening in the test system, which we could measure (Fig. 3). When do we know that a certain measure qualifies as biomarker? In the end, as a result of a validation that shows its predictive value. So how do we increase the odds to choose a meaningful endpoint? A good biomarker has a mechanistic foundation. Thus, the better we understand a given hazard, the easier the choice of measurements. Increasingly, we can use information-rich approaches to choose candidate biomarkers, for example, significantly changed genes in transcriptomics analysis in response to reference toxicants. This will likely result in finding signatures rather than individual genes. The underlying mechanism could make sense of these signatures and separate the signal from the noise. Mapping the underlying pathways will therefore be an important step (see below Ref.[47]). As mechanisms need to translate between model systems (or they cannot be termed models of each other), this is the ideal basis for extrapolation (Fig. 4).
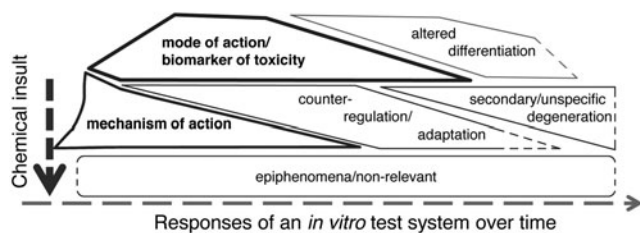

FIG. 3. The different types of responses of an *in vitro* system; reproduced from Kleensang et al. (2016)[35] with permission.
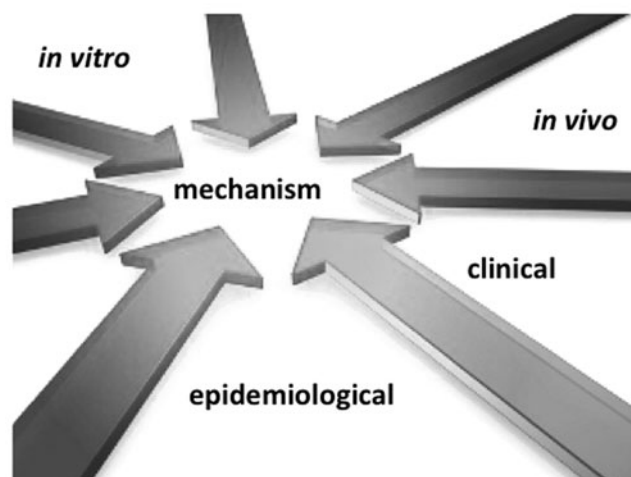

FIG. 4. A common mechanism is crosslinking different models and levels of complexity when studying a toxicological effect.

PoT as Basis of Extrapolation

Toxicity mechanisms may be described by a number of terms, including mode of action, toxicity pathways, AOP, and PoT. The differences are largely academic, with a trend in this sequence from mode of action to PoT to become more molecularly defined and quantitative. The level of resolution increases from current phenomenological assays to Mode of Action, Toxicity Pathway/AOP and Molecular PoT, and ultimately perturbed molecular networks. The terms can, however, be used largely interchangeably.

AOP is the framework developed in the context of OECD. About 200 AOP are currently accepted or under review with many more under development for inclusion in an AOP-Wiki database.[63] These AOP are mostly narrative with a low level of detail (not molecularly defined, not quantitative, no flux, no dynamics); naturally, they are biased by existing knowledge and there is currently no concept for validation. There is some link with the Effectopedia database,[64] which aims for more quantitative pathway information for bioinformatics applications. Increasingly, AOP are used to justify grouping and read-across of chemicals.[65–67]

The term PoT has been coined to describe molecular pathways in the context of the Human Toxome Project.[68–70] PoT are derived by the integration of multiomics approaches and thus have a molecular, high level of detail. The process applies untargeted identification with validation, establishing causality and aiming for quantitative relations and fluxes. The concept of Mechanistic Validation in EBT has been proposed,[71] but could be equally applied to AOP.

The Human Toxome Project and its use of omics thus complement the initiatives of the various U.S. agencies (ToxCast of EPA and Tox21 alliance of National Institutes of Health [NIH], Environmental Protection Agency [EPA], and Food and Drug Administration [FDA], Fig. 5) designed to pragmatically ascertain broad biological characterization of known toxicants by robotized testing.[72–74] While the latter does not primarily identify pathways, it delivers a broad,
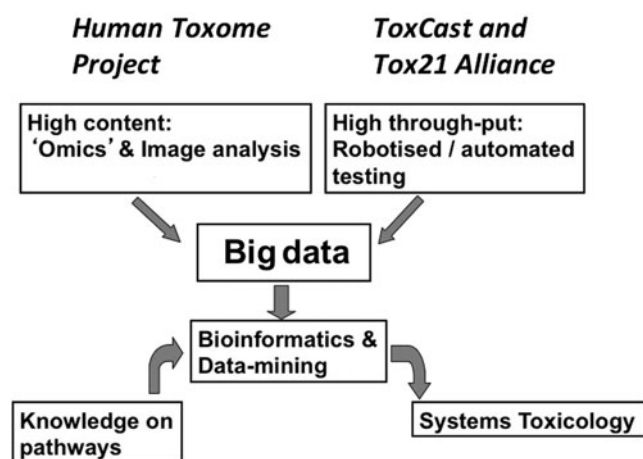
FIG. 5. Twenty-first century technologies creating an information-rich situation, which can be interpreted using pathway knowledge and might ultimately allow establishing a systems toxicology approach.

quality-controlled database for mining and mapping biological activities to hazards and pathways. Notably, high-content imaging is only beginning to contribute to this field.[75]

For PoT identification within the Human Toxome Project, (pre-)validated, robust cell systems are exposed to reference toxicants they are known to correctly predict. Ideally, these have received some regulatory acceptance (available or in progress), and reference substances as well as thresholds of adversity have been defined. Homeostasis under stress,[76] that is, signatures of toxicity established after treatment, is then characterized by orthogonal omics technologies and emerging bioinformatics approaches. Additional knowledge about critical cell infrastructures and networks from molecular biology and biochemistry aid this process. The initial work, funded by NIH and involving six groups, used the well-established endocrine disruptor test using MCF-7 cells (prevalidated by the U.S. Interagency Coordinating Committee on the Validation of Alternative Methods [ICCVAM]) and an initial set of endocrine-disrupting chemicals selected from a priority list of 53 ICCVAM-identified reference compounds. The responses of MCF-7 human breast cancer cells are being phenotyped by transcriptomics and mass spectroscopy-based metabolomics. The bioinformatic tools for PoT deduction represent a core deliverable of the NIH project. It turned out that this given cell line showed tremendous reproducibility issues,[35] however; and it is noteworthy that the MCF-7 test also failed the parallel international validation steered by ICCVAM for reproducibility issues. One lesson from the project is that *a test does not become better by adding a sophisticated endpoint*. On the contrary, characterization of the test system with various omics technologies showed the changes underlying the variability in test results.

The fundamental problem of all omics approaches is that we have too many variables (genes, metabolites, etc.), small *n* (number of measurements), and often considerable noise in the data. This causes tremendous challenges for validation, as discussed in an ECAVM/ICCVAM workshop earlier for transcriptomics, and the way forward requires rigorous quality control[77] to reduce noise and reduction of dimensionality by mapping to pathways leading to targeted follow-up analy-

sis. Notably, OECD is now, 10 years later, considering the first test guidelines based on transcriptomics, that is, for the GARD[78] and SENS-IS[79] assays for skin sensitization. The combined use of orthogonal omics technologies (such as transcriptomics, proteomics, and metabolomics) with analysis of transcription factors (microRNA, protein phosphorylation, etc.) allows cross-validation of approaches. The different omics technologies, it should be noted, do not have the same levels of standardization and quality assurance; part of the human toxome activities is therefore to promote this especially for metabolomics.[80,81] Identified candidates can then be subjected to confirmation with linguistic searches of the respective scientific literature, as recently exemplified for 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) toxicity.[82] Earlier, we showed how a combination of omics technologies pinpoints the PoT for the MPTP metabolite MPP$^+$.[83] Multiomics integration is a key challenge for mapping the Human Toxome. A number of challenges for quality and standardization of cell systems, omics technologies, and bioinformatics, are being addressed. In parallel, concepts for annotation, validation, and sharing of PoT data, as well as their link to adverse outcomes, are being developed. A reasonably comprehensive public database of PoT, the Human Toxome Knowledgebase, could become a point of reference for toxicological research and regulatory test strategies.

Quality assurance of AOP and PoT is the fundamental challenge. This is especially problematic for AOP, which are primarily based on scientific literature, and we have discussed earlier the problems of nonreproducibility of scientific publications.[15] Two analyses by the pharmaceutical industry were particularly damning: Scientists from Amgen[84] reported "Fifty-three papers were deemed 'landmark' studies …scientific findings were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this was a shocking result." Similarly, a group from Bayer found[85]: "…data from 67 projects … revealed that only in ∼20–25% of the projects were the relevant published data completely in line with our in-house findings… In almost two-thirds of the projects, there were inconsistencies between published data and in-house data that either considerably prolonged the duration of the target validation process or, in most cases, resulted in termination of the projects." This is why the author does not believe in using existing knowledge without systematic review to form a point of reference. The quality assurance and confirmation of AOP and PoT—when not done experimentally—should be based on mechanism and evidence, that is, systematic, objective, and transparent compilation of the available literature.[15]

Evidence Integration 1–ITS

Traditionally, toxicology favors stand-alone tests, but more and more often a systematic combination of several information sources is necessary, for example, when single tests do not cover all possible outcomes of interest (e.g., modes of action), classes of test substances (applicability domains), or severity classes of effect; similarly, when positive test results are rare (low prevalence), leading to excessive false-positive results; or when the definitive test is too demanding with respect to costs, work or animal use, where screening allows prioritization for full testing. Furthermore, tests are combined when the human predictivity of any

test alone is too low or when existing evidence shall be integrated. Finally, kinetic information (ADME [absorption, distribution, metabolism, and excretion]) is integrated to make an *in vivo* extrapolation from *in vitro* data.

ITS lend themselves as solution to these challenges.[86] ITS have been proposed around the turn of the century, and some efforts to establish test guidance for regulations have been made. Despite their obvious potential for revamping regulatory toxicology, however, we still have no general agreement on the composition, validation, and adaptation of ITS. Not very different, Weight-of-Evidence and EBT approaches are based on weighing and combining different evidence streams and types of data.

ITS also promise to integrate pathway-based tests, reflecting an adverse outcome pathway, as suggested in Tox-21c.[87] A recent workshop described the state of the art of ITS[88] and discussed earlier suggestions regarding their definition, systematic combination, and quality assurance.[86] We commissioned a whitepaper by Jaworska and Hoffmann[6] that laid the ground for ITS development for skin sensitization.[89] Notably, while ITS are typically used solely for hazard assessment, in the context of OECD, it has recently been broadened to include aspects of exposure and risk assessment. This expansion is termed *Integrated Approaches to Testing and Assessment* (IATA[76]). OECD defines IATA as follows: "IATA are pragmatic, science-based approaches for chemical hazard characterisation that rely on an integrated analysis of existing information coupled with the generation of new information using testing strategies. IATA follow an iterative approach to answer a defined question in a specific regulatory context, taking into account the acceptable level of uncertainty associated with the decision context. There is a range of IATA—from more flexible, non-formalised judgment based approaches (e.g., grouping and read-across) to more structured, prescriptive, rule based approaches [e.g. Integrated Testing Strategy (ITS)]. IATA can include a combination of methods and can be informed by integrating results from one or many methodological approaches [(Q)SAR, read-across, *in chemico, in vitro, ex vivo, in vivo*] or omic technologies (e.g., toxicogenomics)."[8]

One question is how to integrate the various components of ITS. We can obviously follow a Boolean algebra where the components are integrated in an algorithm of decision points. Alternatively, and arguably more promising, the different information sources may be incorporated into a probabilistic hazard assessment, that is, in which different results combine to inform hazard probability. This does not necessarily mean that all components of ITS have to be executed, as methods to identify the next most valuable test and to estimate the possible gain of information by further testing emerge. Ideally, depending on prior knowledge (structure, chemicophysical properties, *in silico* assessments, read-across to similar substances, prior testing, etc.), we can follow an individual path using the most meaningful tests and ending when substantial gains of information are no longer expected.

Evidence Integration 2–Systems Toxicology

The promise of modeling biological systems has given rise to what is now called "systems biology," which our recent glossary[61] defines as: "Study of the mechanisms underlying complex biological processes as integrated systems of many diverse, interacting components. It involves (1) collection of large sets of experimental data (by high-throughput technologies and/or by mining the literature of reductionist molecular biology and biochemistry); (2) proposal of mathematical models that might account for at least some significant aspects of this data set; (3) accurate computer solutions of the mathematical equations to obtain numerical predictions; and (4) assessment of the quality of the model by comparing numerical simulations with the experimental data." In short, this means using high-content (big data) information to literature systematically and model responses in virtual experiments.

It is tempting to translate this to systems toxicology.[90,91] However, frankly, the current examples are rather limited and many older approaches are now simply lumped under the new buzzword. We will likely need to identify a substantial number of quantitative PoT before integrating this knowledge into the systems biology approaches emerging in other disciplines. Ultimately, we hope to create virtual organs and even entire patients. However, the concept reintroduces a more physiological approach to integrated systems after the predominance of the reductionist molecular approaches of the last decades. The internal/external exposure (ADME) considerations will be of critical importance for such systems modeling.[2]

EBT as a Tool for Quality Control and Validation of Extrapolations

The need for quality assurance for these new approaches to spur their development and implementation has been noted.[87] This quality assurance is necessary for the components (cell model, mechanistic basis, measurements, evidence integration) and for their overall validation. Evidence-based medicine (EBM) has revolutionized clinical medicine over the last three decades, illustrating the advantage of objective, critical, and systematic reviews of current practices as well as formal meta-analysis of data and central deposits of current best evidence for a given medical problem. Toxicology might benefit from a similar rigorous review of traditional approaches and the development of meta-analysis tools as well as a central, quality-controlled information portal.[92] Already in 1993, Neugebauer and Holaday[93] in their book *Handbook of Mediators of Septic Shock* showed that EBM methods can be applied to animal studies and *in vitro* work. With Sebastian Hoffmann and his 2005 PhD thesis "Evidence-based *in vitro* toxicology," we developed initial concepts of an EBT. The EBT Collaboration (EBTC)[94] was created in the United States and Europe in 2011 and 2012, respectively.[95] This collaboration of representatives from all stakeholder groups, with the secretariat run by CAAT, aims to develop tools of EBM for toxicology.

EBM was prompted by the need to handle the flood of information in healthcare and to condense the available evidence in an objective manner, including traditional approaches and new scientific developments of variable quality. More than half a million articles are incorporated into Medline every year (an estimated more than 2 million total in medicine). When querying PubMed for the search term "toxicology" for the last 10 years, gives about 30,000 hits, that is, in a database that covers only a fraction of relevant articles in biomedicine. Instead of individuals identifying best evidence for a specific question, the process makes high-quality reviews available at a central deposit

as a primary resource of high-quality condensed information. This requires agreed quality standards, to improve the reliability of the presented evidence and its integration. This is the key difference between evidence-based and narrative reviews (''eminence-based''): most reviews represent a story told by experts presenting their personal opinions on an often broad issue. They tend to favor their own articles and those that fit the argument of their review.

The systematic review, however, proceeds differently. The sources to be included and selection criteria, which articles to consider and which not, are upfront explicitly defined. Before acquiring the actual articles, the search strategy and study document define the procedure for evidence integration. Ideally, this study plan is peer reviewed to safeguard objective, transparent, and efficient processes. The evidence integration includes as a critical element weighing the quality, strength, and possible bias of individual pieces of evidence and how to summarize them as objectively as possible. The latter often involves meta-analysis, that is, statistical approaches combining results from different studies.

Toxicology has a similar problems—information flooding, coexistence of traditional and modern methodologies, and various biases. It is most difficult to find and summarize relevant information for any given question. This has been succinctly illustrated by Christina Ruden,[96] who showed the divergence in judgment and limitations of analysis for 29 cancer risk assessments carried out for trichloroethylene. Four assessments concluded that the substance was carcinogenic, six said it was not, and 19 were equivocal. The main reason for this divergence was a selection bias in the materials considered, that is, an average reference coverage of only 18%, an average citation coverage of most relevant studies of 80%, an interpretation difference of most relevant studies in 27%, and the lack of documentation of study/data quality in 65% of the assessments.

The similar problems of toxicology and clinical medicine, and especially the similarities between setting a diagnosis in medicine and determination of a hazardous substance,[97] prompted us to suggest that EBM tools could be suitable for toxicology.[98] A major step in creating an EBT movement was the *First International Forum Toward an EBT* in 2007,[99] which developed a declaration and 10 defining characteristics of EBT. A working group then developed a consensus definition some time later. The first major development of EBT was the ToxR-Tool to systematically assign quality scores to existing *in vivo* and *in vitro* studies[100] (available as a download from the ECVAM website). Such evaluation is critical for any meta-analysis and also for programs, such as REACH, which use existing information for notifications.

With the creation of the first chair for EBT at Johns Hopkins in 2009, the EBT concept has been institutionalized for the first time at a major academic institution, with hopes that it will serve a starting point for further developments of an EBT movement. Prompted by a CAAT-hosted conference *21ˢᵗ century Validation for 21ˢᵗ Century Tools* in July 2010,[101] a steering group was formed representing several U.S. agencies, industries, academia, and stakeholder organizations. On occasion of the 50th Society of Toxicology Meeting in Washington, on March 11, 2011, the EBT Collaboration was launched, and a donation allowed CAAT to establish its secretariat. The first conference was held in early 2012[102] and a European equivalent launched as a satellite event to EuroTox the same year. A number of working groups have since addressed the EBT tools and governance, carrying out a systematic review of the zebrafish assay for developmental toxicity while developing general guidance for the increasingly used systematic reviews in toxicology.[103,104]

EBT is increasingly contributing to improvement of validation processes.[105] The introduction of retrospective validation by the use of existing data—as first done for the micronucleus test[106]—brings us close to the principles of systematic review of evidence in EBT. Similarly, we need to assess study quality[107] and define our inclusion/exclusion criteria. Concepts of mechanistic validation[71] are emerging using EBT principles to validate AOP. A major step forward would be utilizing validation to evaluate mechanisms of toxicity instead of simply reproducing high-dose animal data of questionable quality. This has to be complemented by defining the toxicity of a validation test compound not just by a single animal test but also by using all available information—especially human data.[16] At the first EBT conference, the close link between EBT and validation was demonstrated with the development concept for the validation of toxicological high-throughput testing approaches.[108] Altogether, Tox-21c and its implementation activities, including the human toxome and the EBTC, promise a credible approach to revamping regulatory toxicology.

Conclusions and General Remarks on *In Vitro* to *In Vivo* Extrapolation in Toxicology

Some final thoughts for extrapolation in toxicology:

- Trash in, trash out: The quality of the input and modeling determines the results (error propagation).
- Intrapolation is easier than extrapolation: We need anchors (points of reference) on both sides of the extrapolation; this will typically be, on the one hand, substances with well-understood toxicities and, on the other hand, associated human hazard manifestations (diseases). Similarly, in the known parts of the chemical universe, we can establish a ''local validity''[65] for our extrapolations because we have ample experience for similar substances. Thus, extrapolations will be easier for industrial chemicals[109] than, for example, nanomaterials,[110–112] where we do not have a known chemical landscape.
- False positives and precaution cannot be extrapolated: Where our precautionary approach has created false positives, these will impair any extrapolation. Thus, we should as clearly as possible distinguish between scientific evidence (risk assessment) and the precautionary decisions as part of risk management.
- Black swans cannot be predicted: Real major hazards (scandals of unsafe products) are rare; these type of events are termed by Taleb as ''black swans.''[113] Black swan events are defined by the ''triplet: rarity, extreme impact and retrospective (though not prospective) predictability.''[114] Absence of evidence of a hazard is not evidence of its absence. Taleb noted ''What is surprising is not the magnitude of our forecast errors, but our absence of awareness of it.'' We presume higher frequencies of hazards to allow the application of Gaussian-type statistics and reasoning. This only creates the belief that we have covered most of the putative hazards, however, and finding so many harmless or borderline events belittles the real threat. Taleb remarked

''True, our knowledge does grow, but it is threatened by greater increases in confidence, which makes our increase in knowledge at the same time an increase in confusion, ignorance, and conceit.''

- Predictions of no toxicity are mostly right, those of toxicity mostly false: We have shown earlier this impact of the low prevalence of a given hazard among industrial chemicals[97]; this is contrary to the regulatory use of novel technologies that typically accept only positive results.
- Extrapolation requires a computer to take over[115]: How does this meet the comfort zone of regulators and the regulated community? Nonetheless, it gives us the tremendous opportunity to use virtual experiments with experimental verification.
- Mechanism as unifying concept: The higher level of integration represents a reduction of dimensionality and noise in our data.
- The central problem is causation in complex networks: We do not really have the frameworks to show causality in networks.
- Black boxes cannot be extrapolated: The more we know (the higher the resolution) the better we can extrapolate.
- *In vitro* biokinetics is lacking: As long as we do not know how much of a test substance actually interacts with a biological system, we cannot extrapolate to *in vivo*.[2]
- Reverse PBPK is a key opportunity: The emerging toolbox of PBPK needs to be adapted for the needs of QIVIVE.[59]
- Extrapolation from single substances to mixtures is largely impossible: The endless number of combinations (including concentrations and timing) impairs analysis. If there are some opportunities, they occur on the level of signatures of toxicity in high-content phenotyping. It might be easier for the subproblem of substances with the same MoA, but partial agonists pose a problem here.
- It is key how we can bring regulators to have sufficient confidence to use the new tools for regulatory use: This clearly represents the bottleneck for change, that is, how to extend the comfort zone of decision takers. A critical element is the objective demonstration of the shortcomings of current tools to open up for change. Targeted communication and education offer establishing confidence by validation and generation change represents critical steps toward this goal.
- We should use the new technologies beyond regulatory toxicology: Frontloading of toxicity assessments in the pharmaceutical industry and Green Toxicology in the chemical arena[116] offer opportunities for early focus of development of substances with lesser toxicological liabilities without the safety concerns mandating extensive validation studies.

Taken together, extrapolation from *in vitro* to *in vivo* is closer to reality as a more general approach to estimating human hazard. To date, we were only successful where we could take shortcuts, for example, topical toxicity (essentially no kinetics, few cells involved), skin sensitization (relatively linear, well-understood AOP), and, arguably, for some toxicities with clear molecular initiating events (estro-genic endocrine disruption) or key events (mutagenicity). Better cell systems and relevant biomarker measurements led by AOP as input, integration of models either as ITS/IATA or as MPS/human-on-a-chip, as well as systems toxicology approaches, together with quality assurance and validation stimulated by evidence-based methodologies, can provide a path to a human-relevant prediction of toxicity.

Acknowledgments

Author Disclosure Statement

The author is the founder of Organome LLC, which has licensed brain organoid technologies from Johns Hopkins to make them commercially available. He also consults Underwriters Laboratories (UL) on computational toxicology.

References

1. Bale AS, Kenyon E, Flynn TJ, et al. Correlating in vitro data to in vivo findings for risk assessment. ALTEX 2014:31;79–90.
2. Tsaioun K, Blaauboer BJ, Hartung T. Evidence-based absorption, distribution, metabolism, excretion and toxicity (ADMET) and the role of alternative methods. ALTEX 2016:33;343–358.
3. Krewski D, Acosta D, Jr, Andersen M, et al. Toxicity Testing in the 21st Century: A Vision and a Strategy. J Toxicol Environ Health B Crit Rev 2010:13;51–138.
4. Krewski D, Westphal M, Andersen ME, et al. A framework for the next generation of risk science. Environ Health Perspect 2014:122;796–805.
5. Ankley GT, Bennett RS, Erickson RJ, et al. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 2010:29;730–741.
6. Jaworska J, Hoffmann S. Integrated Testing Strategy (ITS)—Opportunities to better use existing data and guide future testing in toxicology. ALTEX 2010:27; 231–242.
7. Combes RD, Balls M. Integrated testing strategies for toxicity employing new and existing technologies. Altern Lab Anim 2011:39;213–225.
8. OECD. 2016. Integrated Approaches to Testing and Assessment (IATA). www.oecd.org/chemicalsafety/riskassessment/iata-integrated-approaches-to-testing-and-assessment.htm (last accessed Dec. 17, 2016).
9. Plant NJ. An introduction to systems toxicology. Toxicol Res 2015:4;9–22.
10. Sturla SJ, Boobis AR, FitzGerald RE, et al. Systems toxicology: From basic research to risk assessment. Chem Res Toxicol 2014:27;314–329.
11. Sauer JM, Hartung T, Leist M, et al. Systems toxicology: The future of risk assessment. Int J Toxicol 2015:34;346–348.
12. Stephens ML, Betts K, Beck NB, et al. The emergence of systematic review in toxicology. Toxicol Sci 2016:152;10–16.
13. Hartung T. Food for thought … on animal tests. ALTEX 2008:25;3–9.

14. Hartung T. (2015) Evolution of toxicological science: the need for change. Int J Risk Assess Manage 2017:20; 21–45.

15. Hartung T. Look back in anger–what clinical studies tell us about preclinical work. ALTEX 2013:30;275–291.

16. Hoffmann S, Edler L, Gardner I, et al. Points of reference in validation—The report and recommendations of ECVAM Workshop. Altern Lab Anim 2008:36;343–352.

17. Hartung T. Food for thought … on validation. ALTEX 2007:24;67–72.

18. Leist M, Hasiwa M, Daneshian M, et al. Validation and quality control of replacement alternatives–current status and future challenges. Toxicol Res 2012:1;8.

19. Leist M, Hartung T. Inflammatory findings on species extrapolations: Humans are definitely no 70-kg mice. Arch Toxicol 2013:87;563–567.

20. Adriaens E, Barroso J, Eskes C, et al. Retrospective analysis of the Draize test for serious eye damage/eye irritation: Importance of understanding the in vivo endpoints under UN GHS/EU CLP for the development and evaluation of in vitro test methods. Arch Toxicol 2014:88;701–723.

21. Luechtefeld T, Maertens A, Russo DP, et al. Analysis of draize eye irritation testing and its prediction by mining publicly available 2008–2014 REACH data. ALTEX 2016: 33;123–134.

22. Hoffmann S. LLNA variability: An essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. ALTEX 2015:32;379–383.

23. Luechtefeld T, Maertens A, Russo DP, et al. Analysis of publically available skin sensitization data from REACH registrations 2008–2014. ALTEX 2016:33;135–148.

24. Dumont C, Barroso J, Matys I, et al. Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. Toxicol In Vitro 2016:34;220–228.

25. Smirnova L, Harris G, Leist M, et al. Cellular resilience. ALTEX 2015:32;247–260.

26. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell 4th ed. New York: Garland Science; 2002. www.bioon.com/book/biology/mboc/mboc.cgi@code= 220801800040279.htm (last accessed Dec. 18, 2016).

27. Boobis AR, Cohen SM, Dellarco V, et al. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Crit Rev Toxicol 2006:36;781–796.

28. Boobis AR, Doe JE, Heinrich-Hirsch B, et al. IPCS framework for analyzing the relevance of a noncancer mode of action for humans. Crit Rev Toxicol 2008:38; 87–96.

29. Rossini GP, Hartung T. Towards tailored assays for cell-based approaches to toxicity testing. ALTEX 2012:29;359–372.

30. Hartung T. Food for thought … on cell culture. ALTEX 2007:24;143–147.

31. Hartung T, Daston G. Are in vitro tests suitable for regulatory use? Toxicol Sci 2009:111;233–237.

32. Olson H, Betton G, Robinson D, et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. Regul Toxicol Pharmacol 2000:32;56–67.

33. Pound P, Ebrahim S, Sandercock P, et al. Where is the evidence that animal research benefits humans? BMJ 2004: 328;514–517.

34. Coecke S, Ahr H, Blaauboer BJ, et al. Metabolism: A bottleneck in in vitro toxicological test development. Altern Lab Anim 2006:34;49–84.

35. Kleensang A, Vantangoli M, Odwin-DaCosta S, et al. Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. Sci Rep 2016:6;28994.

36. Coecke S, Balls M, Bowe G, et al. Guidance on good cell culture practice.Altern Lab Anim 2005:33;261–287.

37. Stacey GN, Hartung T. Availability, standardization and safety of human cells and tissues for drug screening and testing. In: Drug Testing In Vitro: Breakthroughs and Trends in Cell Culture Technology. U Marx, V Sandig (eds); pp. 231–250. Weinheim: WILEY-VCH Verlag; 2007.

38. Pamies D, Bal-Price A, Simeonov A, et al. Good cell culture practice for stem cells and stem-cell-derived models. ALTEX 2017:34;95–132.

39. OECD, 2016 Draft guidance document on Good In Vitro Method Practices (GIVIMP) for the development of in vitro methods for regulatory use in human safety assessment. www.oecd.org/env/ehs/testing/OECD_Draft_GIVIMP_ in_Human_Safety_Assessment.pdf (last accessed Dec. 14, 2016).

40. Hartung T. 3D—A new dimension of in vitro research. Advanced Drug Delivery Reviews, Preface Special Issue ''Innovative tissue models for in vitro drug development,'' 2014:69;vi.

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

42. Hartung T, Leist M. Food for thought … on the evolution of toxicology and phasing out of animal testing. ALTEX 2008:25;91–96.

43. Leist M, Hartung T, Nicotera P. The dawning of a new age of toxicology. ALTEX 2008:25;103–114.

44. Hartung T. Toxicology for the twenty-first century. Nature 2009:460;208–212.

45. Hartung T. From alternative methods to a new toxicology. Eur J Pharm Biopharm 2011:77;338–349.

46. Willett C, Caverly Rae J, Goyak KO, et al. Building shared experience to advance practical application of pathway-based toxicology: Liver toxicity mode-of-action. ALTEX 2014:31;500–519.

47. Hartung T, McBride M. Food for thought… on mapping the human toxome. ALTEX 2011:28;83–93.

48. Andersen M, Betts K, Dragan Y, et al. Developing microphysiological systems for use as regulatory tools—challenges and opportunities. ALTEX 2014:31;364–367. Extended online version www.altex.ch/resources/altex_2014_3_Suppl_ Andersen.pdf (last accessed June 11, 2017).

49. Hartung T, Zurlo J. Alternative approaches for medical countermeasures to biological and chemical terrorism and warfare. ALTEX 2012:29;251–260.

50. NRC–National Research Council, Committee on Animal Models for Assessing Countermeasures to Bioterrorism Agents. Animal Models for Assessing Countermeasures to Bioterrorism Agents. Washington, DC: The National Academies Press; 2011, pp. 1–153. http://dels.national academies.org/Report/Animal-Models-Assessing-Counter measures/13233 (last accessed June 11, 2017).

51. Hogberg HT, Bressler J, Christian KM, et al. Toward a 3D model of human brain development for studying gene/ environment interactions. Stem Cell Res Ther 2013:4 (Suppl 1) S4;1–7.

52. Pamies D, Hartung T, Hogberg HT. Biological and medical applications of a brain-on-a-chip. Exp Biol Med 2014: 239;1096–1107.

53. Pamies D, Barreras P, Block K, et al. A Human Brain Microphysiological System derived from iPSC to study central nervous system toxicity and disease. ALTEX 2017 [Epub ahead of print]. doi: 10.14573/altex.1609122. www.altex.ch/resources/epub_Pamies2_of_161128_v2.pdf (last accessed Dec. 13, 2016).

54. Smirnova L, Hogberg HT, Leist M, et al. Developmental neurotoxicity–challenges in the 21st century and in vitro opportunities. ALTEX 2014:31;129–156.

55. Marx U, Andersson TB, Bahinski A, et al. Biology-inspired microphysiological system approaches to solve the prediction dilemma of substance testing using animals. ALTEX 2016:33;272–321.

56. Smirnova L, Harris G, Delp J, et al. A LUHMES 3D dopaminergic neuronal model for neurotoxicity testing allowing long-term exposure and cellular resilience analysis. Arch Toxicol 2016:90;2725–2743.

57. Tsaioun K, Blaauboer BJ, Hartung T. Evidence-based absorption, distribution, metabolism, excretion and toxicity (ADMET) and the role of alternative methods. ALTEX 2016:33;343–358.

58. Bouvier d'Yvoire M, Prieto P, Blaauboer BJ, et al. Physiologically-based Kinetic Modelling (PBK Modelling): Meeting the 3Rs Agenda. The Report and Recommendations of ECVAM Workshop 63. Altern Lab Anim 2007:35;661–671.

59. Basketter DA, Clewell H, Kimber I, et al. A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. ALTEX 2012:29;3–89.

60. Leist M, Hasiwa N, Rovida C, et al. Consensus report on the future of animal-free systemic toxicity testing. ALTEX 2014:31;341–356.

61. Ferrario D, Hartung T. Glossary of reference terms for alternative test methods and their validation. ALTEX 2014:31; 319–335.

62. Blaauboer BJ, Boekelheide K, Clewell HJ, et al. The use of biomarkers of toxicity for integrating in vitro hazard estimates into risk assessment for humans. ALTEX 2012:29;411–425.

63. Adverse Outcome Pathway Wiki. https://aopwiki.org (last accessed Dec. 19, 2016).

64. Effectopedia. www.qsari.org/index.php/software/100-effectopedia (last accessed Dec. 19, 2016).

65. Patlewicz G, Ball N, Becker RA, et al. Read-across approaches—Misconceptions, promises and challenges ahead. ALTEX 2014:31;387–396.

66. Hartung T. Making big sense from big data in toxicology by read-across. ALTEX 2016:33;83–93.

67. Ball N, Cronin MTD, Shen J, et al. Toward Good Read-Across Practice (GRAP) guidance. ALTEX 2016:33;149–166.

68. Human Toxome Project. http://humantoxome.com (last accessed Dec. 19, 2016).

69. Bouhifd M, Hogberg HT, Kleensang A, et al. Mapping the Human Toxome by systems toxicology. Basic Clin Pharm Toxicol 2014:115;1–8.

70. Kleensang A, Maertens A, Rosenberg M, et al. Pathways of toxicity. ALTEX 2014:31;53–61.

71. Hartung T, Stephens M, Hoffmann S. Mechanistic validation. ALTEX 2013:30;119–130.

72. Hartung T, Stephens M. Toxicity testing in the 21st Century, approaches to implementation. In: *Encyclopedia of Toxicology* 3rd ed. P Wexler (ed); pp. 673–675. London: Elsevier Inc., Academic Press.

73. Hartung T. Lessons learned from alternative methods and their validation for a new toxicology in the 21st century. J Toxicol Environ Health 2010:13;277–290.

74. Davis M, Boekelheide K, Boverhof DR, et al. The new revolution in toxicology: The good, the bad, and the ugly. Ann N Y Acad Sci 2013:1278;11–24.

75. van Vliet E, Danesian M, Beilmann M, et al. Current approaches and future role of high content imaging in safety sciences and drug discovery. ALTEX 2014:31; 479–493.

76. Tollefsen KE, Scholz S, Cronin MT, et al. Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). Regul Toxicol Pharmacol 2015:70;629–640.

77. Corvi R, Ahr H-J, Albertini S, et al. Validation of toxicogenomics-based test systems: ECVAM-ICCVAM/ NICEATM considerations for regulatory use. Environ Health Perspect 2006:114;420–429.

78. Johansson H, Albrekt A-S, Borrebaeck CAK, et al. The GARD assay for assessment of chemical skin sensitizers. Toxicol In Vitro 2013:27;1163–1169.

79. Cottrez F, Boitel E, Auriault C, et al. Genes specifically modulated in sensitized skins allow the detection of sensitizers in a reconstructed human skin model. Development of the SENS-IS assay. Toxicol In Vitro 2015:29; 787–802.

80. Bouhifd M, Hartung T, Hogberg HT, et al. Review: Toxicometabolomics. J Appl Toxicol 2013:33;1365–1383.

81. Ramirez T, Daneshian M, Kamp H, et al. Metabolomics in toxicology and preclinical research. ALTEX 2013:30; 209–225.

82. Maertens A, Luechtefeld T, Kleensang A, et al. MPTP's pathway of toxicity indicates central role of transcription factor SP1. Arch Toxicol 2015:89;743–755.

83. Krug AK, Gutbier S, Zhao L, et al. Transcriptional and metabolic adaptation of human neurons to the mitochondrial toxicant MPP+. Cell Death Dis 2014:5;e1222.

84. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012:483;531–533.

85. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? Nat Rev Drug Disc 2011:10;712–712.

86. Hartung T, Luechtefeld T, Maertens A, et al. Integrated testing strategies for safety assessments. ALTEX 2013:30;3–18.

87. Hartung T. A toxicology for the 21st century: Mapping the road ahead. Toxicol Sci 2009:109;18–23.

88. Rovida C, Alépée N, Api AM, et al. Integrated testing strategies (ITS) for safety assessment. ALTEX 2015:32; 25–40.

89. Jaworska J, Harol A, Kern PS, et al. Integrating non-animal test information into an adaptive testing strategy - skin sensitization proof of concept case. ALTEX 2011:28;211–225.

90. Hartung T, van Vliet E, Jaworska J, et al. Systems toxicology. ALTEX 2012:29;119–128.

91. Hartung T, FitzGerald R, Paul J, Miriams G, Peitsch M, Rostami-Hodjegan A, Shah I, Wilks M, and Sturla S. Systems toxicology - Real world applications and opportunities. Chemical Research in Toxicology 2017:30; 870–882.

92. Hartung T. Food for thought… on evidence-based toxicology. ALTEX 2009:26;75–82.

93. Neugebauer EA, Holaday JW. *Handbook of Mediators in Septic Shock* 1ˢᵗ ed. Boca Raton, Florida: CRC-Press; 1993, pp. 1–608.

94. www.ebtox.org (last accessed Dec. 19, 2016).

95. Hoffmann S, Stephens M, Hartung T. Evidence-based toxicology. In: *Encyclopedia of Toxicology* 3ʳᵈ ed. vol 2. P Wexler (ed); pp. 565–567. London: Elsevier Inc., Academic Press; 2014.

96. Ruden C. The use and evaluation of primary data in 29 trichloroethylene carcinogen risk assessments. Regul Toxicol Pharmacol 2001:34;3–16.

97. Hoffmann S, Hartung T. Diagnosis: Toxic!–Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. Toxicol Sci 2005:85;422–428.

98. Hoffmann S, Hartung T. Towards an evidence-based toxicology. Hum Exp Toxicol 2006:25;497–513.

99. Griesinger C, Hoffmann S, Kinsner-Ovaskainen A, et al. Proceedings of the first international forum towards evidence-based toxicology. Conference Centre Spazio Villa Erba, Como, Italy. 15–18 October 2007. Human and Experimental Toxicology, Special Issue: Evidence-Based Toxicology (EBT) 2009:28;81–163.

100. Schneider K, Schwarz M, Burkholder I, et al. ''ToxRTool'', a new tool to assess the reliability of toxicological data. Toxicol Lett 2009:189;138–144.

101. Hartung T. Evidence based-toxicology–the toolbox of validation for the 21st century? ALTEX 2010:27;241–251.

102. Stephens ML, Andersen M, Becker RA, et al. Evidence-based toxicology for the 21st century: Opportunities and challenges. ALTEX 2013:30;74–104.

103. Stephens ML, Betts K, Beck NB, et al. (2016). The emergence of systematic review in toxicology. Toxicol Sci 2016:152:10–16.

104. Hoffmann S, de Vries RBM, Stephens ML, et al. A primer on systematic reviews in toxicology. Archives of Toxicology 2017, in press. DOI 10.1007/s00204-017-1980-3, available online: https://link.springer.com/article/10.1007/s00204-017-1980-3.

105. Hartung T, Bremer S, Casati S, et al. A modular approach to the ECVAM principles on test validity. Altern Lab Anim 2004:32;467–472.

106. Corvi R, Albertini S, Hartung T, et al. ECVAM retrospective validation of *in vitro* micronucleus test (MNT). Mutagenesis 2008:23;271–283.

107. Samuel GO, Hoffmann S, Wright R, et al. Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. Environ Int 2016:92–93;630–646.

108. Judson R, Kavlock R, Martin M, et al. Perspectives on validation of high-throughput pathway-based assays supporting the 21st century toxicity testing vision. ALTEX 2013:30;51–66.

109. Hartung T. Food for thought… on alternative methods for chemical safety testing. ALTEX 2010:27;3–14.

110. Hartung T. Food for thought… on alternative methods for nanoparticle safety testing. ALTEX 2010:27;87–95.

111. Hartung T, Sabbioni E. Alternative *in vitro* assays in nanomaterial toxicology. WIREs Nanomed Nanobiotechnol 2011:3;545–573.

112. Silbergeld EK, Contreras EQ, Hartung T, et al. Nanotoxicology: ''the end of the beginning''–Signs on the roadmap to a strategy for assuring the safe application and use of nanomaterials. ALTEX 2011:28;236–241.

113. Bottini AA, Hartungm T. Food for thought… on economics of animal testing. ALTEX 2009:26;3–16.

114. Taleb NN. *The black swan–the impact of the highly improbable*. New York: The Random House Publishing Group; 2007.

115. Hartung T, Hoffmann S. Food for thought on…. in silico methods in toxicology. ALTEX 2009:26;155–166.

116. Maertens A, Anastas N, Spencer PJ, et al. Green Toxicology. ALTEX 2014:31;243–249.

Address correspondence to:
Thomas Hartung, MD, PhD
Center for Alternatives to Animal Testing
Bloomberg School of Public Health
Johns Hopkins University
615 N. Wolfe Street
Baltimore, MD 21205

E-mail: thartun1@jhu.edu