Food for Thought ...
Mechanistic Validation

Thomas Hartung [1,2], Sebastian Hoffmann [2,3], and Martin Stephens [1]

[1]Johns Hopkins Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, USA;
[2]University of Konstanz, CAAT-Europe, Germany; [3]seh consulting, Paderborn, Germany

Summary

Validation of new approaches in regulatory toxicology is commonly defined as the independent assessment of the reproducibility and relevance (the scientific basis and predictive capacity) of a test for a particular purpose. In large ring trials, the emphasis to date has been mainly on reproducibility and predictive capacity (comparison to the traditional test) with less attention given to the scientific or mechanistic basis. Assessing predictive capacity is difficult for novel approaches (which are based on mechanism), such as pathways of toxicity or the complex networks within the organism (systems toxicology). This is highly relevant for implementing Toxicology for the 21st Century, either by high-throughput testing in the ToxCast/ Tox21 project or omics-based testing in the Human Toxome Project. This article explores the mostly neglected assessment of a test's scientific basis, which moves mechanism and causality to the foreground when validating/qualifying tests. Such mechanistic validation faces the problem of establishing causality in complex systems. However, pragmatic adaptations of the Bradford Hill criteria, as well as bioinformatic tools, are emerging. As critical infrastructures of the organism are perturbed by a toxic mechanism we argue that by focusing on the target of toxicity and its vulnerability, in addition to the way it is perturbed, we can anchor the identification of the mechanism and its verification.
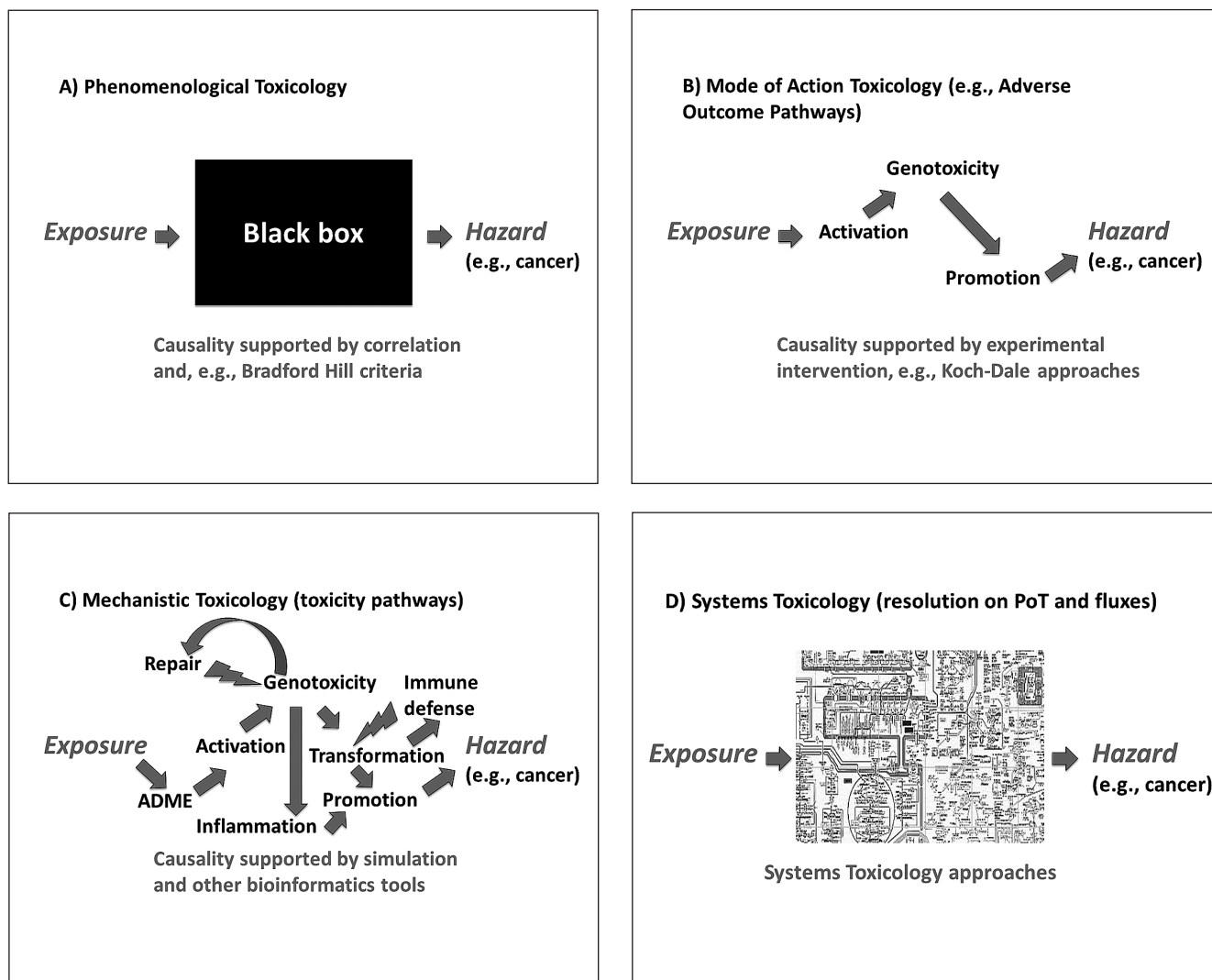
Keywords: regulatory toxicology, Tox-21c, validation, alternatives to animal testing, systems biology

Introduction

This series of articles offers perspectives on areas requiring change, mainly in (regulatory) toxicology such as the assessment of chemicals (Hartung, 2010c), cosmetics (Hartung, 2008b), food (Hartung and Koëter, 2008), medical countermeasures (Hartung and Zurlo, 2012), nanoparticles (Hartung, 2010b), and earlier drugs (Hartung, 2001), as well as basic research (Gruber and Hartung, 2004). The shortcomings of current approaches using animals (Hartung, 2008a), cells (Hartung, 2007b), or *in silico* methods (Hartung and Hoffmann, 2009) have been discussed. In line with the roadmap for alternatives to animal-based systemic toxicity testing (Basketter et al., 2012), integrated testing strategies (Hartung et al., 2013) and pathway of toxicity (PoT)-based approaches (Hartung and McBride, 2011; Hartung et al., 2012) were presented. As shown in Figure 1, this follows a change in paradigm from phenomenological toxicology (Fig. 1A) to mode-of-action-based toxicology (Fig. 1B), to mechanistic toxicology (Fig. 1C), and finally to systems toxicology (Fig. 1D). The change from (c) to (d) illustrates the transition

from systems structure to systems dynamics. In a simple traffic analogy: At the first (phenomenological) level, we understand that our car (model) drove from city A (exposure) to city B (hazard manifestation), but we do not know which route it took. At the mode of action level, we understand the route. At the next (mechanistic) level, we see the complexity of interfering events. At the systems level, we model the dynamics of fluxes, roadblocks, deviations, counter-regulatory events, etc.

The opportunities and needs for quality assurance have already been discussed twice in this series of articles (Hartung, 2007a, 2009) as well as in a publication of our Transatlantic Think Tank for Toxicology (t[4]) (Hartung, 2010a) and (Leist et al., 2012). Very often we touched on the need for a mechanistic approach to testing that generates relevant evidence, which can then be compiled to inform decision-making. In this paper, we address this mechanistic thinking with respect to the problem of confirming a biological mechanism and using established mechanisms as the basis for validating our test systems. Thus, it is a discussion of biological causality in a field that is increas-

A) Phenomenological Toxicology

Exposure → **Black box** → *Hazard* (e.g., cancer)

Causality supported by correlation and, e.g., Bradford Hill criteria

B) Mode of Action Toxicology (e.g., Adverse Outcome Pathways)

Genotoxicity

Exposure → **Activation** → **Promotion** → *Hazard* (e.g., cancer)

Causality supported by experimental intervention, e.g., Koch-Dale approaches

C) Mechanistic Toxicology (toxicity pathways)

Repair
Genotoxicity **Immune defense**
Exposure **Activation** **Transformation** *Hazard* (e.g., cancer)
ADME **Promotion**
Inflammation

Causality supported by simulation and other bioinformatics tools

D) Systems Toxicology (resolution on PoT and fluxes)

Exposure → → *Hazard* (e.g., cancer)

Systems Toxicology approaches

Fig. 1: The evolution of toxicology from (A) phenomenology to (B), mode of action to (C), mechanism to (D), systems approaches

ingly becoming aware of the complexity of the organism and embracing a systems toxicology approach. We present several aspects that we consider essential when embarking on mechanistic validation.

The classical definition of validation was coined in 1990 at an ECVAM/ERGATT workshop (Balls et al., 1990):

> "Validation is the process by which the reliability and relevance of a new method is established for a specific purpose."

Later redefinitions of the process (OECD, 2005) were more detailed:

> "Test method validation is a process based on scientifically sound principles … by which the reliability and relevance of a particular test, approach, method, or process are established for a specific purpose. Reliability is defined as the extent of reproducibility of results from a test within and among laboratories over time, when performed using the same standardised protocol. The relevance of a test

method describes the relationship between the test and the effect in the target species and whether the test method is meaningful and useful for a defined purpose, with the limitations identified. In brief, it is the extent to which the test method correctly measures or predicts the (biological) effect of interest, as appropriate. Regulatory need, usefulness, and limitations of the test method are aspects of its relevance. New and updated test methods need to be both reliable and relevant, i.e., validated."

The importance of the scientific basis was proposed by Worth and Balls (2001). The modular approach (Hartung et al., 2004), a consensus between ECVAM and ICCVAM, introduced this aspect of scientific validity and referred also to the prediction model:

> "Validation is a process in which the scientific basis and reproducibility of a test system, and the predictive capacity of an associated prediction model, undergo independent assessment."

While the modular approach made it into the OECD guidance document on validation, it is quite remarkable that this definition was not embraced. The challenges to the current validation paradigm, such as the imperfections of the reference test, the inability to demonstrate that a new test is better than the reference test, the costs and duration of the current process, and its failure – to date – to be adopted to testing strategies, have been discussed elsewhere (Hartung, 2007a; Leist et al., 2012). In addition, we have earlier stressed the opportunity that lies in this aspect of scientific basis (Hartung, 2010a; Hartung and Zurlo, 2012).

Consideration 1:
Validation of mechanism or mechanistic validation?

Biomedical science addresses how living organisms work and how proper functioning can be disturbed or restored. When moving to a systems approach, this is all about mechanism, i.e., a level of resolution lower than the macroscopic and phenomenological view. It is about the "How?" Toxicology has embraced a focus on mechanism for a couple of decades and we have termed it "mechanistic," "predictive," "translational," etc. Some, when fearing that the promise to identify the mechanism might be difficult to realize in practice, introduced "mode of action" to allow for uncertainty in characterizing the mechanism. As defined in the US EPA draft, *Mechanisms and Mode of Dioxin Action*[1], mechanism of action is "*the detailed molecular description of key events in the induction of cancer or other health endpoints*," whereas mode of action refers to "*the description of key events and processes, starting with interaction of an agent with the cell through functional and anatomical changes, resulting in cancer or other health endpoints*."

"Research has to be hypothesis-driven" is the fundamental approach – almost a mantra – in biomedical sciences. Such research typically corresponds to suggesting a mechanism and then using a specific ''known'' example to demonstrate it. This approach has its shortcomings, especially, as noted by Popper (1963), science can only falsify a hypothesis, because:

1. *"It is easy to obtain confirmations, or verifications, for nearly every theory – if we look for confirmations.*
2. *Confirmations should count only if they are the result of risky predictions; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory – an event which would have refuted the theory.*
3. *Every "good" scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is.*
4. *A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think), but a vice.*
5. *Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability; but there are degrees of testability: some theories are more testable, more exposed to refutation, than others; they take, as it were, greater risks.*
6. *Confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory. (I now speak in such cases of "corroborating evidence.")*
7. *Some genuinely testable theories, when found to be false, are still upheld by their admirers – for example, by introducing ad hoc some auxiliary assumption, or by reinterpreting the theory ad hoc in such a way that it escapes refutation. Such a procedure is always possible, but it rescues the theory from refutation only at the price of destroying, or at least lowering, its scientific status….)*

One can sum up all this by saying that the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability."

The difficulty of scientific work is that we have to verify our hypothesis of causality, i.e., mechanism. Once we have deduced it, we cannot just aim to destroy it. In the same way, we cannot select hypotheses that are unconditionally destroyed or altered. So we need frameworks of "*corroborating evidence*" (Popper, above) to come as close as possible to proving the hypothesis – in our case, causality. The classical frameworks of Koch-Dale and Bradford Hill were already discussed in the last article in this series (Hartung et al., 2013). They take somewhat different approaches, as they originate from different centuries (i.e., before and after Popper). Koch's postulates were aimed at giving unambiguous proof of causality for a pathogen causing a disease. When translated to physiology by Dale, the idea remained to request similar evidence as for pathogenesis of an infectious disease, which together makes the case of a linear causality of mediation of an effect. The problem is that few things in biology are linear and networked systems are too complex to provide certainty when interrogated, given that most experiments only remain valid if some variables are kept constant. Sir Bradford Hill (Hill, 1965), in contrast, gave a number of types and pieces of evidence that support causality without the assumption of a simple linear relationship. It is undoubtedly the more adequate framework for complex systems, in his case epidemiology, and, thus, for a systems toxicology approach.

The beauty of the Koch-Dale approach lies in its straightforward guidance on which experiments to carry out to determine causality. It asks for a mediator (originally a disease agent; in Koch's case a microbial pathogen): Show that the mediator is present when the disease state forms and show that you can protect the organism by blocking its formation or action and that you can induce (or aggravate) the disease state by its (co-) application. Translated to the paradigm of *Toxicity Testing in the 21st Century: A Vision and a Strategy* (NRC, 2007) or Tox-21c, for a pathway of toxicity (PoT), this means: show it, block it and induce it. If these experiments agree, we are on a good track to confirming the PoT.

---

[1] Available at http://www.epa.gov/ncea/pdfs/dioxin/nas-review/pdfs/part3/dioxin_pt3_ch03_oct2004.pdf

Interestingly, Hackney and Linn (2013) reformulated Koch's postulates for environmental toxicology as:

"(1) a definable environmental chemical agent must be plausibly associated with a particular observable health effect; (2) the environmental agent must be available in the laboratory in a form that permits realistic and ethically acceptable exposure studies to be conducted; (3) laboratory exposures to realistic concentrations of the agent must be associated with effects comparable to those observable in real-life exposures; (4) the preceding findings must be confirmed in at least one investigation independent of the original."

This reformulation, however, is relatively weak in reference to causality ("plausibly associated") and stresses only the reproducibly induced effects in an experimental model. Just as there are many ways to Rome, there are many ways to hazard manifestation. Plausibility is not proof. The "*confirmation in at least one investigation independent of the original*" is also quite questionable, especially as counter-evidence is not mentioned.

More recently, Adami (2011), suggested combining Bradford Hill criteria and elements of evidence-based toxicology, named an EPID-TOX approach. Recognizing the difficulty of aligning toxicological and epidemiological data, they stress the uncertainty of results and aim to give guidance how to move out of uncertainty to a positive or negative association. The approach adds more to the field of data integration than to causation.

Validating a mechanism in toxicology means establishing the causality between toxicant and hazard manifestation and identification of how it happens. Together the two approaches (Koch/Dale and Bradford Hill) help to support (not prove) causality, but only by establishing causality between toxicant and hazard. They can be used for confirming a mechanism when applied to the mediating events. This means that, in principle, for each and every event of a PoT we need to establish causality. Neither framework was developed for causality in toxicology and Bradford Hill was very careful to offer his criteria as a comparative standard, i.e., it is only valid if there is no better plausible alternative explanation of the effect. In our case, the comparative standard would be the scientific evidence supporting a specific mechanism. In order to maximize existing knowledge and minimize subjectivity in establishing standards, a central, frequently updated repository of accumulated mechanistic knowledge is required.

Notably, there is no institution for collecting the evidence for a certain mechanism to be responsible for causing an effect, nor is there a repository for retrieving the information once accumulated. This is exactly what the Human Toxome Project (Hartung and McBride, 2011; Baker, 2013) attempts for toxicology, which admittedly is only a small part of the life sciences. It is based on the notion that groups of toxicants leading to similar hazard manifestations likely employ the same or similar mechanisms (pathways of toxicity), resulting in the same disturbed physiology. An alternative view might be that there are only a certain number of meta-stable physiological states a disturbed biology can assume before collapsing, and they are linked with some probability to particular hazard manifestations. If we have

a mobile (in homeostatic conditions) and we cut one element off it, the mobile will either collapse (die) or assume another meta-stable state, but it cannot assume an endless number of different states (Fig. 2). It might thus be simpler and more helpful to describe these states as signatures of toxicity (SoT) rather than pathways of toxicity (PoT).

This is very much consistent with the thrust of a recent paper by Liu et al. (2013) on the observability of complex systems. They state: "*Although the simultaneous measurement of all internal variables, like all metabolite concentrations in a cell, offers a complete description of a system's state, in practice experimental access is limited to only a subset of variables, or sensors. A system is called observable if we can reconstruct the system's complete internal state from its outputs. … We apply this approach to biochemical reaction systems, finding that the identified sensors are not only necessary but also sufficient for observability.*" It will be most interesting to see whether this is applicable to our problem, i.e., the description of the cellular state after toxicant exposure by measuring a variety but not all metabolites (or gene expressions). In Liu et al.'s case, about 10% of the influential nodes were sufficient to describe the state of the system.

In conclusion, validating the mechanism of a (group of) toxicant(s) is the basis for mechanistic validation of tests that identify those toxicants.

Consideration 2: Ascertaining mechanism and causality in complex systems

Describing a complex system is not the same as confirming a mechanism within it. The criteria of Bradford Hill (see Box 1) make an association more likely (probably sufficient for validation purposes), but they are tailored more to simple, linear associations (though the field of epidemiology from which these criteria originate has to handle highly complex systems).


Box 1

Bradford Hill criteria (Hill, 1965)

– Strength: The stronger an association between cause and effect the more likely a causal interpretation, but a small association does not mean that there is not a causal effect.

– Consistency: Consistent findings of different persons in different places with different samples increase the causal role of a factor and its effect.

– Specificity: The more specific an association is between factor and effect, the bigger the probability of a causal relationship.

– Temporality: The effect has to occur after the cause.

– Biological gradient: Greater exposure should lead to greater incidence of the effect with the exception that it can also be inverse, meaning greater exposure leads to lower incidence of the effect.
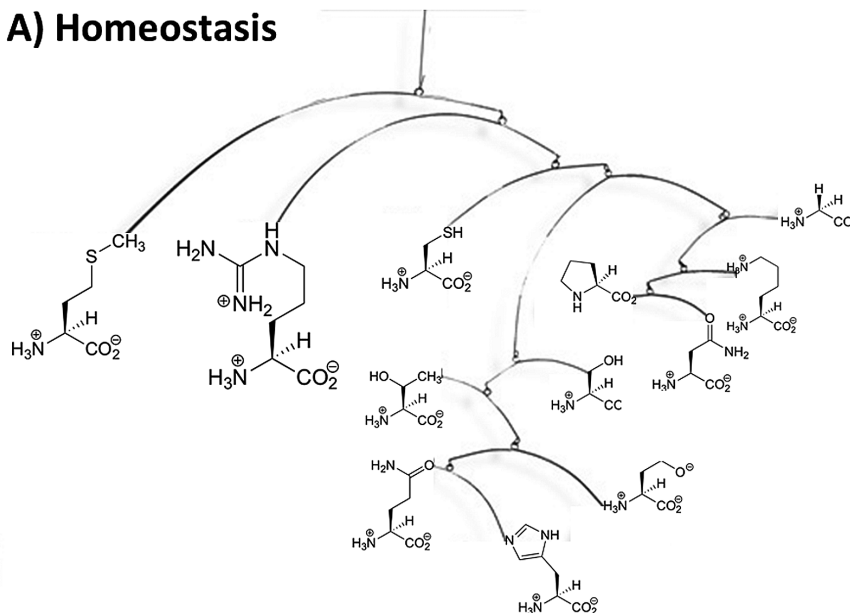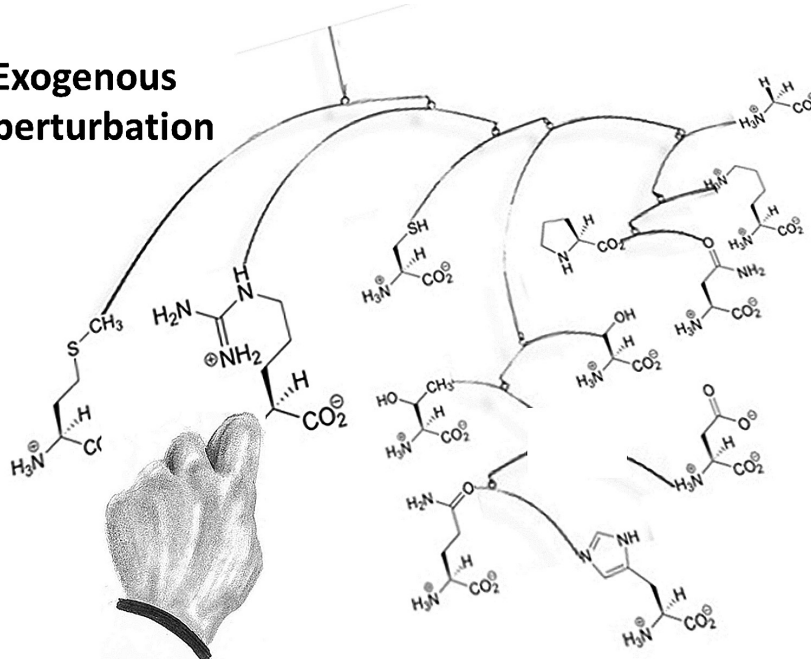
Fig. 2: Illustration how the homeostasis (here depicted by a mobile of some amino acids) is perturbed and new homeostasis under stress forms (the metabolomics signature of this perturbation), which is meta-stable as the system rearranges if the stress is discontinued

– Plausibility: A possible mechanism between factor and effect increases the causal relationship, with the limitation that knowledge of the mechanism is limited by best available current knowledge.
– Coherence: A coherence between epidemiological and laboratory findings leads to an increase in the likelihood of this effect. However, the lack of laboratory evidence cannot nullify the epidemiological effect on the associations.
– Experiment: Similar factors that lead to similar effects increase the causal relationship of factor and effect.

Modifications of the Bradford Hill criteria by Susser (1991) stress "predictive performance," which he defines deductively as "*the ability of a causal hypothesis drawn from an observed association to predict an unknown fact that is consequent on the initial association*." This is reminiscent of traditional validation, where test accuracy, i.e., the proportion of correct results when challenged with a new set of reference compounds, is the key for declaring validity. The aspect of mechanism is covered by Bradford Hill or Susser under "coherence," e.g., Susser:

"*5. Coherence is defined by the extent to which a hypothesized causal association is compatible with preexisting theory and knowledge. Coherence can be considered in terms of many subclasses.*

5.1 Theoretical coherence requires compatibility with preexisting theory.

*5.2 Factual coherence requires compatibility with preexisting knowledge.*

5.3 Biologic coherence requires compatibility with current biologic knowledge that is drawn from species other than human or, in humans, from levels of organization other than the unit of observation, especially those less complex than the person.

5.4 Statistical coherence requires compatibility with a comprehensible or, at the least, conceivable model of the distribution of cause and effect (it is enhanced by simple distributions readily comprehended – for instance, a dose-response relation – and is obscured by those that are nonlinear and complex)."

By *mechanistic validation* we have something slightly different in mind (Fig. 3), i.e., moving away from correlation of phenomena toward the molecular description of pathways (Hartung and McBride, 2011). Put simply, the steps that should be part of mechanistic validation are:

– Condense the knowledge of biological/mechanistic circuitry (in the absence of xenobiotic challenge) underlying the hazard in question
– Compile evidence that reference chemicals leading to the hazard in question perturb the biology in question, i.e., mainly pathway identification by using reference substances in valid(ated) models and experimental proof of their role
– Develop a test that purports to reflect this biology
– Verify that toxicants shown to employ this mechanism also do so in the model
– Verify that interference with this mechanism hinders positive test results

This still proves mediation at every step, but with plausibility and the respective experimental underpinning. First, we would show that a certain mechanism is involved and whether it is nec-
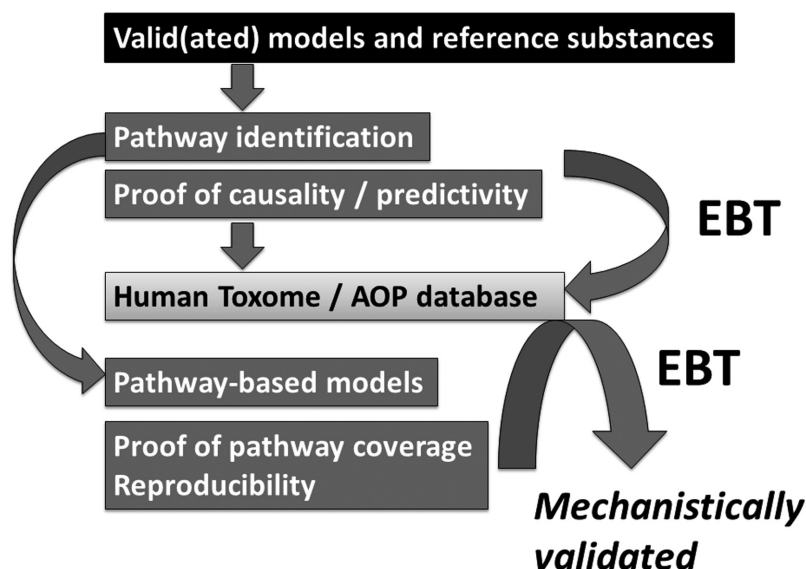
Fig. 3: The Mechanistic Validation Scheme for test systems with a possible role for Evidence-based Toxicology (EBT) type of assessments

essary and/or sufficient or aggravating. Then we can whether a given test reflects this mechanism. In contrast to traditional validation, this will not require testing of large numbers of new substances. Rather, it entails identifying toxicants that result in the same hazard in question and showing that they employ the same mechanism as the chemical used to deduce the PoT in the pathway-based test. We should keep in mind that, unlike epidemiology, where the conceptual frameworks by Bradford Hill and Susser originate, toxicology can typically use experimental interventions, though with all the limitations of using models as discussed earlier (Hartung, 2007b, 2008a).

A key question is: how should we assess a chemical lacking hazard information in the absence of mechanistic information? Can we use the following information to test a chemical whose mechanism of action is unknown? We will need (1) knowledge of biological/mechanistic circuitry relevant to xenobiotic challenge, (2) tests that purport to reflect key mechanisms in biology, and (3), verification that toxicants that have been shown to employ one or more of these mechanisms also do so in the test system. This might be done even in a relatively small part of the chemical universe; we have termed this approach "test-across" (similar to read-across) (Hartung, 2007a), i.e., creating local applicability domains by showing that (structurally) related substances are correctly identified.

In the end, it will have to be shown whether we need to echo Douglas Adams (in his book *Dirk Gently's Holistic Detective Agency*): "*The complexities of cause and effect defy analysis.*"

Consideration 3:
Complexity – Bioinformatics taking over…

The types of causality analysis considered so far are suited to relatively simple forms of causation. Chemicals can obviously act through multiple mechanisms. Identification of one mechanism does not necessarily imply that it is the only, or even the key, mechanism. Whether this type of reasoning helps us "change the world, one PoT at a time" needs to be shown. We should be aware, however, of the emerging bioinformatics toolbox for exploring the network structures in large, complex datasets, especially Granger causality and dynamic Bayesian network interference (Zou and Feng, 2009). The previous paper in this series discussed a new approach to causation originating in ecological modeling (Sugihara et al., 2012). Whether this offers an avenue to systematically test causality in large datasets from omics and/or high-throughput testing, needs to be explored.

However, it is worth considering the nature and the consequences of studying complex systems. Another approach is an interesting analysis from the military[2], which aims to "*take a very pragmatic approach to causality as the production and propagation of effects*":

"*What makes systems complex is the network of interdependencies between the elements of the system. This means that consequences of any event or property*

develop through many interacting pathways, and similarly, that if we want to know how a particular property or event came about, we will find that there were many cross-linked pathways that contributed to it. In such a system therefore, we cannot expect simple causality (one cause – one effect), or linear causal chains …to hold in general….
It is possible to make a complex system appear simpler by restricting the scope of attention to a particular pathway, but if the scope is widened to include other pathways, or if unexpected side-effects that have propagated through those pathways are linked back and suddenly manifested within the restricted scope, we are quickly reminded that the causal chain was just one of many pathways through a network…. Such systems are therefore inherently characterised not by linear causal chains, but by networks of causal relationships through which consequences propagate and interact. Such networks of interactions between contributing factors can exhibit emergent behaviours which are not readily attributable or comprehensible."

This complexity has posed formidable challenges to our ability to characterize these phenomena completely. However, the report not only argues against "*inappropriately linear causal thinking in complex systems*" (by the way the hallmark of hypothesis-driven research) but also identifies several helpful approaches: (1) Bayesian techniques (Korb and Nicholson, 2003), (2) Systems dynamics (Sterman, 2001), (3) Network theory (Newman et al., 2006), (4) Simulation, especially agent-based simulation (Epstein, 1999), and (5) Non-linear dynamical systems (Katok and Hasselblatt, 1997). These are all machine-learning tools, i.e., we have to hand things over to Captain Computer.

However, we have one advantage in toxicology compared to other complex systems, such as society, military conflicts, financial markets: If the organism does not drop dead, it has to develop a new homeostasis under stress (Hartung et al., 2012). Only certain meta-stable conditions can be assumed by the organism (Fig. 2), which correspond with the typical signatures of toxicity and which can be observed with omics technologies. Such states are described in the report[2] as attractors, i.e., "*regions of the possibility space of the situation which are more likely to be occupied than other regions, and in the extreme, once entered, are very difficult to escape from… In complex adaptive systems, what creates the attractor are the dynamic adaptive processes of agents acting in their own interests. This creates a dynamic stability as opposed to an intrinsic lowest-energy stability.*" This suggests exploiting such states of lower uncertainty to describe the state of the complex system and possibly the transition from the prior (normal) state.

One challenge is determining the extent to which this complexity needs to be reflected in the complexity of the cell models used for PoT identification. The necessary quality control of validated *in vitro* systems, especially longer-term 3D systems once in routine use (Hartung and Zurlo, 2012), is enormous. A

---

[2] The Technical Cooperation Program (2010). Causal & influence networks in complex systems.
Available at: http://www.lifelong.ed.ac.uk/

Good Cell Culture Practice (Coecke et al., 2005) for complex *in vitro* systems has yet to be developed (Leist et al., 2010). Calls for quality assurance in complex models are increasing, e.g., LeCluyse et al. (2012): "*In the future, data generated from studies utilizing* in vitro *organotypic model systems should be judged or scrutinized in light of a system's ability to maintain or exhibit certain biochemical properties at physiologic levels, not just as the presence or absence of key functions or components, which so often occurs today in published reports.*"

Another obvious challenge is posed by chemicals with mechanisms that are a series of necessary steps. An example would be botulinum neurotoxin, where cleavage of the relevant cellular proteins precludes neurotransmission but prior steps are critical (internalization of the neurotoxin into the cell, transport within the cell, etc.). The full process is akin to a molecular adverse outcome pathway (AOP). Such complex phenomena are difficult to discern from a complex dataset with sophisticated software. It requires the system to be broken down into the individual steps and then simulated, framing the untargeted analysis of PoT identification. One might argue that this is actually moving away from truly untargeted analysis, as complexity is reduced to certain windows of interest. But should a mechanistic test capture all such steps? Is it enough to just screen for a final step? This likely depends, in part, on regulatory context/purpose (e.g., batch testing versus lot release in our neurotoxin example).

Russell and Burch (1959) discussed the relative merits of fidelity and discrimination models. They distinguished between high-fidelity models, such as rodents and other laboratory mammals in toxicity testing (used because of their general physiological and pharmacological similarity to humans), and high discrimination models that "*reproduce one particular property of the original, in which we happen to be interested.*" They warned of the "*high-fidelity fallacy*" and of the danger of expecting discrimination in particular circumstances from models that show high fidelity in other, more general terms. Zurlo et al. (1996) refer to other more recent analyses of the differing molecular responses to certain chemicals by the rat, the mouse, and the human: "*Russell and Burch pointed out that the fidelity of mammals as models for man is greatly overestimated; however, replacement alternatives methods must be based on good science, and extravagant claims that cannot be substantiated must be avoided.*" We must be careful not to uncritically produce new high fidelity models, but our complex simulations are prone to exactly this as they model the past and give the impression therefore also to cover the future, the prediction of new effects. Assays should be based on the lowest level of biological/biochemical organization that still demonstrates the mechanism; a pertinent example would be pH readings versus the Draize eye test. Another example, discussed by Russell and Burch (cited above), is Niko Tinbergen's representation of a mother gull by a red dot on a fake beak, which by itself elicited the appropriate food-begging behavior from chicks (pecking at the red dot to elicit food regurgitation). In the end,

we need red-dot recognizing systems, not a complex reconstruction of the entire organism to react properly, to make the right decisions regarding toxicity.

Consideration 4:
Causation versus method evaluation – the fusion of two roots of evidence-based toxicology

We have stressed (Hartung, 2009) that the call for evidence-based toxicology (EBT) has two roots – Philip Guzelian's group's proposal for a more rigorous approach to causation of chemical effects (Guzelian et al., 2005), and ours (Hoffmann and Hartung, 2006) on seeking new approaches to method evaluation. The proposal for a mechanistic validation fuses the two concepts and uses causation to evaluate methods. By ascertaining mechanistic validity (Hartung, 2010a; Hartung and Zurlo, 2012) we can qualify/assess (avoiding the term "validate," which is typically used for the correlative traditional validation approaches) both the components of ITS (Hartung et al., 2013) and high-throughput tests (Judson et al., 2013).

Guzelian et al. suggest the following to establish a cause-and-effect relationship:

> *"Having assembled and critically evaluated the 'knowledge,' how do we decide if the evidence permits an evidence-based conclusion of general causation? For experimental data, the matter seems reasonably straightforward. The results of well conducted RCTs [randomized controlled trials], like controlled laboratory experiments with animal or* in vitro *systems that exhibit strength (statistical), specificity, temporality, dose-dependence and predictive performance especially if replicated (consistency) and supported by mechanism/ pathophysiology (biologic plausibility, coherence), lead to an evidence-based conclusion of cause and effect (i.e., the establishment of a risk)."*

If it was only that "straightforward"…

It is instructive to recall the problem of the cancer bioassay for carcinogenicity (Basketter et al., 2012), though we might say that the assay simply lacks specificity and predictive performance. However, this brings us back to point zero – the need to build better tests and to identify and verify the mechanisms involved and to provide quantitative data for them.

Our earlier use of the term "qualification" (of a test), borrows from FDA's approaches (Goodsaid and Frueh, 2007): "*The pharmacogenomics guidance[3] defines a valid biomarker as 'a biomarker that is measured in an analytical test system with well-established performance characteristics and for which there is an established scientific framework or body of evidence that elucidates the physiologic, toxicologic, pharmacologic, or clinical significance of the test results.' The validity of a biomarker is closely linked to what we think we can do with it. This biomarker context drives not only how we define a biomarker but also*

---

[3] US Food and Drug Administration. Guidance for industry – pharmacogenomic data submissions. Available at: www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM079849.pdf Accessed March 28, 2013.

the complexity of its qualification." This adds to the performance characteristics the notions of significance and usefulness ("*what we think we can do with it*"). It seems fair to translate "significance" to mechanistic relevance. The aspect of "usefulness" adds a restriction to a given area of application (similar to the applicability domain for a test, which we introduced with the modular approach to validation (Hartung et al., 2004) and to some extent the expectation that relevant predictions are made in this realm. The notion of usefulness apparently lessens expectations about explicit predictions of the results of a reference test. We earlier stressed that the main similarity of Evidence-based Medicine and EBT is actually clearer when viewing a toxicological method as a diagnostic test (Hoffmann and Hartung, 2005; Hartung, 2010a). It is interesting that this discussion has been largely driven by test accuracy and very little by mechanism, which is quite different to biomarker qualification.

By suggesting EBT as a starting point for method validation for Tox-21c (Hartung, 2010a) and thus for mechanistic methods, we are facilitating convergence on the basis of causation. The first outcome was a whitepaper on the validation of high-throughput methods (Judson et al., 2013) as used in ToxCast in the context of the first North American EBT conference (2012) (Stephens et al., 2013). The next logical step is establishing the mechanistic basis of assays used in the HTS. This is a tremendous opportunity for the EBT Collaboration (http://www.ebtox.com).

EBT incorporates, from its role model Evidence-based Medicine, the overarching evidence-based principles of transparency, objectivity, and consistency. These defining characteristics assist any process, whether based on mechanism or correlation, in surviving peer scrutiny. EBT offers more than the actual result of a systematic review and creates the possibility of continuous improvement in the light of additional evidence. A high-quality assessment of the state of the evidence will always also be an assessment of the uncertainty and the limitations of the data. This, by itself, is as valuable as the actual condensation of the available evidence.

Consideration 5:
The point of reference for mechanistic validation

Validations of new methods have traditionally been carried out by comparing them to the tests they aim to replace, with the problematic assumption that pre-existing tests represent a gold standard. As the results of the reference test are classifications, the classified toxicants are the point of reference. An important ECVAM workshop discussing points of reference for validation (Hoffmann et al., 2008) suggested a move to a composite point of reference, where all knowledge of toxicants is used to create the correct classification. This allows, for example, sorting false-positive and -negative results. The goal is no longer to reproduce the traditional test with all its shortcomings but to define what an ideal test would identify.

How does this change if we make mechanism the central criterion? John Frazier first suggested using mechanism for validation (Frazier, 1994) but there was no follow-up. If we now

move in this direction (Hartung, 2007a), we will have to build a consensus on a relevant mechanism and its contribution to hazard manifestation. The Human Toxome Project aims to develop the process for doing exactly this. To put it simply: no agreed mechanism, no mechanistic validation. The Human Toxome Project does not aim to confirm known/presumed pathways, but to be open to new causal links. We would quickly run out of pathways if we focused only on those already known. We also would only reinforce our biases, overstressing what we *believe* to know compared to what we want to know. For this reason, the project begins with untargeted analyses of chemically induced metabolites and transcripts. By associating the patterns of change (i.e., the signatures of toxicity (SoT)) to pathways (the pathways of toxicity (PoT)), the noise common to all systems is eliminated. The two orthogonal technologies, as well as replicates and concentration/response relationships around the thresholds of adversity, further focus PoT identification.

It is important to keep in mind that such a mechanistic validation does not necessarily need reference chemicals, nor does it rely on animal data as gold standards. In principle, it wants to facilitate the shift to human biology under Tox-21c – for this purpose, the validation can rely, for example, on the use of a cell or tissue's own biochemicals (agonists, antagonists, enzymes, hormones, etc.) to show biological relevance of the pathway in the test system, besides the use of known xenobiotic disrupters (toxicants, pharmacological as well as scientific inhibitors such as antibodies and silencing RNAs) of a mechanism, to show merit of the assay.

Consideration 6:
Simulation as virtual experiment to challenge the consistency of mechanism and our understanding of the complex system

The good news of a systems toxicology approach to safety assessments is that it is gaining human relevance; the bad news is that human reality is complex (Kitano, 2002). The interactions of tens of thousands of genes, millions of gene products, and thousands of metabolites are far beyond our comprehension. And if we are somehow capable of modeling such a system by reducing it to its nodes and other key components, our assessment should no longer be based on comparisons to models of similar complexity (animals) and their results. The opportunity lies in modeling outcomes (Hartung et al., 2012) and verifying/optimizing the models in comparison to the human data. Emerging examples from the Virtual Embryo Project of US EPA best illustrate this approach – for example, feeding test data into models and comparing the models' reaction to *in vivo* responses (Knudsen and DeWoskin, 2011; Knudsen et al., 2013).

This approach is quite different to modeling future events (see our comments in Bottini and Hartung, 2009; Hartung and Hoffmann, 2009) – here we discuss models based on the input of experimental data and the cross-validation of models' predictions by experiments. This has little to do with the processes of forecasting critically discussed in books like *The black swan – the impact of the highly improbable* (Taleb, 2007) and

Useless arithmetic – why environmental scientists can't predict the future (Pilkey and Pilkey-Jarvis, 2007), as the biological simulation represents more a sequence of virtual and real experiments informing each other.

Consideration 7:
Vulnerability and critical cellular infrastructure – a different look at the same problem?

There is a certain similarity between a toxic insult to an organism and a (natural or manmade) hazard to a society. Whether the outcome of perturbation leads to a societal disaster or a toxicity hazard depends in both cases on exposure and vulnerability. The vulnerability perspective has become common in the field of societal disasters, where one of the coauthors had responsibilities in the past[4]. It is tempting to translate some of these concepts to toxicology. What is the relation to causality? It is changing the focus from what is affecting to what is affected. If we study toxicants, we often see many perturbations, but might we be able to sort out those which are particularly meaningful because they harm an Achilles' heel of the cell? This might narrow down our identification of causative pathways, which we need to confirm.

Vulnerability has been defined (Radvanovsky, 2006) as "*an inherent weakness in a system or its operating environment that may be exploited to cause harm to the system*" or as physical vulnerability (Starr, 1969) "*essentially related to the degree of exposure and the fragility of the exposed elements in the action of the phenomena*." An alternative definition comes from Wisner et al. (2005): "*By vulnerability we mean the characteristics of a person or group and their situation that influence their capacity to anticipate, cope with, resist and recover from the impact of a natural hazard (an extreme natural event or process).*" This brings us closer to the organism view and includes defense and repair, i.e., resilience. We will come back to this in a moment.

The vulnerability perspective has become very common in hazard and disaster studies. The concept of Critical Infrastructures (Radvanovsky, 2006) is key for understanding and mapping vulnerabilities, defined as "*assets of physical and computer-based systems that are essential to the minimum operations of the economy and the government.*" In toxicology, this might be translated as *structures, functions and information flows that are essential for the minimum operations of the organism and its decision making.*

This leads to a slightly different risk concept, and it is important to keep the different terms separate: "*In the same way that for many years the term risk was used to refer to what is today called hazard, currently, many references are made to the word vulnerability as if it were the same thing as risk. It is important to emphasize that these are two different concepts…*" (O. D. Cardona in Bankoff et al., 2004).

This might add an interesting component to the toxicological paradigm: The dose makes the poison, but the individual makes the disease. Or more technically:

Risk = Exposure x Hazard x Vulnerability

Cardona remarks (Bankoff et al., 2004): "*Risk is a complex and, at the same time, curious concept. It represents something unreal, related to random chance and possibility, with something that still has not happened. It is imaginary, difficult to grasp and can never exist in the present, only in the future. If there is certainty, there is no risk.*" Along these lines, risk exists because of our uncertainty in exposure, hazard, vulnerability, and the associations between them.

Traditional toxicologists will likely state that vulnerability is already part of the risk assessment process, especially where vulnerable subpopulations (children and the elderly, for example) are considered. It is important that one should also consider vulnerability on a cellular level. What are the Achilles' heels of the cell? We might focus on pathways leading to perturbations, for which the cell has little redundancy and repair capacity, and which are critical for cell survival and functionality as well as the overall function of the organ and organism.

Conclusions

Mechanistic thinking opens new avenues for assessing the performance of test methods. Such thinking bases our confidence not on correlation (the number of storks declining with the number of births in many countries) but on the accumulated knowledge of how a particular exposure leads to particular effects. This approach requires certainty in our deduction of mechanism and becomes more difficult as we acknowledge the complexity of systems and our lack of understanding thereof. If we assume that causation is linear, we have a simple approach to prove it (Koch-Dale). If we take complexity into account we are left with ascertaining a relationship (Bradford Hill). As we increase our understanding of the system we are studying we can begin to model and carry out virtual experiments to understand causality and verify these predictions by experiments.

This opens up the possibility of a mechanistic validation, especially where the type of information generated does not directly correspond to a high-quality point of reference. This approach entails the danger that it is based on our current level of understanding: for example, before identifying *Helicobacter pylori* as causative agent, stress-induced hypersecretion of hydrochloric acid was considered the main cause of gastritis, ulcers, and stomach cancer. When scientific paradigms change, we have to review what we concluded from the old concepts, but it might still be better to base our regulatory science on the current understanding of pathophysiology and not on pure correlations.

What does this mean for the validation process? The key change will be the introduction of a module for scientific rel-

---

[4] TH headed the Traceability, Risk and Vulnerability Assessment Unit of the Institute for the Protection and Security of the Citizen, EU Joint Research Centre, Ispra, Italy.

evance into the 7-step modular approach (Hartung et al., 2004). We do not suggest making this a new module 8 (scientific relevance) but rather to add it as a new option to existing module 5 (predictive relevance). The latter would become module 5a, with scientific relevance becoming module 5b. As stressed earlier (Judson et al., 2013), for high-throughput methods it will be necessary to compensate for often lacking information on inter-laboratory reproducibility (module 3), as often no adequate facilities for ring trials are available, but within-laboratory variability is low anyway. Again we might consider that strengthening our assessment with mechanistic relevance might help here, though it provides a different type of confirmation. It might be promising to start formally validating the mechanistic basis of assays in the current large scale high-throughput testing programs in toxicology (ToxCast and Tox-21 project).

The obvious practical problem with Mechanistic Validation is that it depends on our current understanding of the system and the identified mechanisms. Some might argue that we need full understanding of the system, which we can never attain. However, being aware that we can only approximate (model) the system, we can test the predictivity for some, but not all areas, where we do have a point reference. Deduction and annotation of mechanisms are key prerequisites for a Mechanistic Validation. Creating such a repository, or knowledge base, of pathways of toxicity (PoT) is the goal of the Human Toxome Project. Although its governance has not been established, consensus on the process and types of information to be compiled is emerging. A t^4 (Transatlantic Think Tank for Toxicology) workshop on this topic was held in Baltimore in October, 2012 and the report is underway.

The EBT toolbox lends itself to a Mechanistic Validation as it offers processes to compile and evaluate evidence objectively and transparently (Hartung, 2010a). It might become the sparring partner for new method development and quality assurance. However, it might as well be conceived that the traditional validation process could embrace the same approaches. The fact that both ECVAM (Hartung, 2010d) and ICCVAM (Birnbaum, 2013) are currently undergoing redefinition offers such opportunities to tackle the challenge of validation of 21st century technologies.

References

Adami, H.-O., Berry, S. C. L., Breckenridge, C. B., et al. (2011). Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. *Toxicol Sci 122*, 223-234.

Baker, M. (2013). The 'omes puzzle. *Nature 494*, 416-419.

Balls, M., Blaauboer, B., Brusick, D., et al. (1990). Report and recommendations of the CAAT/ERGATT workshop on the validation of toxicity testing procedures. *Altern Lab Anim 18*, 313-336.

Bankoff, G., Frerks, G., and Hilhorst, D. (2004). *Mapping vulnerability: Disasters, Development and People*. Sterlin VA: Earthscan, pp. 236.

Basketter, D. A., Clewell, H., Kimber, I., et al. (2012). A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing – t^4 report. *ALTEX 29*, 3-91.

Birnbaum, L. S. (2013). 15 years out: Reinventing ICCVAM. *Environ Health Perspect 121*, a40.

Bottini, A. A. and Hartung, T. (2009). Food for thought ... on the economics of animal testing. *ALTEX 26*, 3-16.

Coecke, S., Balls, M., Bowe, G., et al. (2005). Guidance on good cell culture practice. a report of the second ECVAM task force on good cell culture practice. *Altern Lab Anim 33*, 261-287.

Epstein, J. M. (1999). Generative social science: Studies in agent-based computational modeling. *Complexity 4*, 41-60.

Frazier, J. M. (1994). The role of mechanistic toxicology in test method validation. *Toxicol In Vitro 8*, 787-791.

Goodsaid, F. and Frueh, F. (2007). Biomarker qualification pilot process at the US Food and Drug Administration. *AAPS 9*, E105-E108.

Gruber, F. P. and Hartung, T. (2004). Alternatives to animal experimentation in basic research. *ALTEX 21, Suppl 1*, 3-31.

Guzelian, P. S., Victoroff, M. S., Halmes, N. C., et al. (2005). Evidence-based toxicology: a comprehensive framework for causation. *Human Exp Toxicol 24*, 161-201.

Hackney, J. D. and Linn, W. S. (2013). Koch's postulates updated: a potentially useful application to laboratory research and policy analysis in environmental toxicology. *Am Rev Respir Dis 119*, 849-852.

Hartung, T. (2001). Three Rs potential in the development and quality control of pharmaceuticals. *ALTEX 18, Suppl 1*, 3-13.

Hartung, T., Bremer, S., Casati, S., et al. (2004). A modular approach to the ECVAM principles on test validity. *Altern Lab Anim 32*, 467-472.

Hartung, T. (2007a). Food for thought ... on validation. *ALTEX 24*, 67-80.

Hartung, T. (2007b). Food for thought ... on cell culture. *ALTEX 24*, 143-152.

Hartung, T. (2008a). Food for thought … on animal tests. *ALTEX 25*, 3-9.

Hartung, T. (2008b). Food for thought ... on alternative methods for cosmetics safety testing. *ALTEX 25*, 147-162.

Hartung, T. and Koëter, H. (2008). Food for thought ... on food safety testing. *ALTEX 25*, 259-264.

Hartung, T. (2009). Food for thought ... on evidence-based toxicology. *ALTEX 26*, 75-82.

Hartung, T. and Hoffmann, S. (2009). Food for thought ... on in silico methods in toxicology. *ALTEX 26*, 155-166.

Hartung, T. (2010a). Evidence-based toxicology – the toolbox of validation for the 21st century? *ALTEX 27*, 253-263.

Hartung, T. (2010b). Food for thought ... on alternative methods for nanoparticle safety testing. *ALTEX 27*, 87-95.

Hartung, T. (2010c). Food for thought ... on alternative methods for chemical safety testing. *ALTEX 27*, 3-14.

Hartung, T. (2010d). Comparative analysis of the revised Directive 2010/63/EU for the protection of laboratory animals with its predecessor 86/609/EEC – a t^4 report. *ALTEX 27*, 285-303.

Hartung, T. and McBride, M. (2011). Food for thought ... on mapping the human toxome. *ALTEX 28*, 83-93.

Hartung, T. and Zurlo, J. (2012). Food for thought ... Alternative approaches for medical countermeasures to biological and chemical terrorism and warfare. *ALTEX 29*, 251-260.

Hartung, T., van Vliet, E., Jaworska, J., et al. (2012). Food for thought ... Systems toxicology. *ALTEX 29*, 119-128.

Hartung, T., Luechtefeld, T., Maertens, A., et al. (2013). Integrated testing strategies for safety assessments. *ALTEX 30*, 3-18.

Hill, A. B. (1965). The environment and disease: association or causation? *Proc R Soc Med 58*, 295-300.

Hoffmann, S. and Hartung, T. (2005). Diagnosis: toxic! – trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations. *Toxicol Sci 85*, 422-428.

Hoffman, S. and Hartung, T. (2006). Toward an evidence-based toxicology. *Hum Exp Toxicol 25*, 497-513.

Hoffmann, S., Edler, L., Gardner, I., et al. (2008). Points of reference in the validation process: the report and recommendations of ECVAM Workshop 66. *Altern Lab Anim 36*, 343-352.

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

Katok, A. and Hasselblatt, B. (1997). *Introduction to the modern theory of dynamical systems*. Cambridge University Press, pp. 802.

Kitano, H. (2002). Systems biology: a brief overview. *Science 295*, 1662-1664.

Knudsen, T. B. and DeWoskin, R. S. (2011). Systems modeling in developmental toxicity. In *General, Applied and Systems Toxicology*. Wiley Online Library. doi: 10.1002/9780470744307.gat236

Knudsen, T., Martin, M., Chandler, K., et al. (2013). Predictive models and computational toxicology. *Meth Molec Biol 947*, 343-374.

Korb, K. B. and Nicholson, A. E. (2003). *Bayesian Artificial Intelligence*. Chapman and Hall/CRC. pp. 392.

LeCluyse, E. L. E., Witek, R. P. R., Andersen, M. E. M., et al. (2012). Organotypic liver culture models: meeting current challenges in toxicity testing. *Critic Rev Toxicol 42*, 501-548.

Leist, M., Efremova, L., and Karreman, C. (2010). Food for thought ... considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317.

Leist, M., Hasiwa, N., Daneshian, M., and Hartung, T. (2012). Validation and quality control of replacement alternatives – current status and future challenges. *Toxicol Res 1*, 8-22.

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2013). Observability of complex systems. Proc Natl Acad Sci USA 110, 2460-2465.

Newman, M., Barabási, A.-L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton, USA: Princeton University Press, pp. 582.

NRC – National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and A Strategy*. Washington, D.C.: National Academy Press.

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. ENV/JM/MONO(2005)14.

Pilkey, O. H. and Pilkey-Jarvis, L. (2007). *Useless arithmetic – why environmental scientists can't predict the future*. New York, USA: Columbia University Press.

Popper, K. R. (1963). Science as Falsification. In: *Conjectures and Refutations*. London: Routledge and Keagan Paul, pp. 33-39.

Radvanovsky, R. (2006). *Critical Infrastructure: Homeland Security and Emergency Preparedness*. Boca Raton, USA: CRC Taylor and Francis, pp. 302.

Russell, W. M. S. and Burch, R. L. (1959). *The principles of humane experimental technique*. London, UK: Methuen. pp. 238. http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc.

Starr, C. C. (1969). Social benefit versus technological risk. *Science 165*, 1232-1238.

Stephens, M. L., Andersen, M., Becker, R. A., et al. (2013). Evidence-based toxicology for the 21st century: Opportunities and challenges. *ALTEX 30*, 74-104.

Sterman, J. D. (2001). System dynamics modeling. *California Manag Rev 43*, 8-25.

Sugihara, G., May, R., Ye, H., et al. (2012). Detecting causality in complex ecosystems. *Science 338*, 496-500.

Susser, M. M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am J Epidemiol 133*, 635-648.

Taleb, N. N. (2007). *The black swan – the impact of the highly improbable*. New York, USA: The Random House Publishing Group.

Wisner, B., Blaikie, P., Cannon, T., and Routledge, I. D. (2005). *At risk – natural hazards, people's vulnerability and disasters*. 2nd edition. London: Routledge. 471 pp.

Worth, A. P. and Balls, M. (2001). The importance of the prediction model in the validation of alternative tests. Alternatives to laboratory animals. *Altern Lab Anim 26*, 135-144.

Zou, C. and Feng, J. (2009). Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics 10*, 122.

Zurlo, J. J., Rudacille, D. D., and Goldberg, A. M. A. (1996). The three Rs: the way forward. *Environ Health Persp 104*, 878-880.

Acknowledgements

Correspondence to

Thomas Hartung, MD PhD
Center for Alternatives to Animal Testing
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
W7032, Baltimore, MD 21205, USA
e-mail: thartung@jhsph.edu