Concept Article

Uncertainties of Testing Methods: What Do We (Want to) Know About Carcinogenicity?

Martin Paparella [1], Annamaria Colacci [2] and Miriam N. Jacobs [3]

[1]Chemicals & Biocides, Environment Agency Austria, Vienna, Austria; [2]Agency for Prevention, Environment and Energy, Emilia-Romagna, Italy; [3]Department of Toxicology, Centre for Radiation, Chemical and Environmental Hazards Public Health England, Chilton, Oxfordshire, UK

Summary

An approach to systematically describe the uncertainties and complexity of the standard animal testing and assessment approach for carcinogenicity is explored by using a OECD Guidance Document that was originally developed for reporting defined *in vitro* approaches to testing and assessment. The format is suitable for this re-purposing and it appears that the potential multitude of approaches for integrating and interpreting data from standard animal testing may ultimately be conceptually similar to the challenge of integrating relevant *in vitro* and *in silico* data. This structured approach shall allow 1) fostering interest in developing improved defined *in silico* and *in vitro* approaches; 2) the definition of what type of effects should be predicted by the new approach; 3) selection of the most suitable reference data and assessments; 4) definition of the weight that the standard animal reference data should have compared to human reference data and mechanistic information in the context of assessing the fitness of the new *in vitro* and *in silico* approach; 5) definition of a benchmark for the minimum performance of the new approach, based on a conceptual recognition that correlation of alternative assessment results with reference animal results is limited by the uncertainties and complexity of the latter. A longer term perspective is indicated for evolving the definition of adversity for classification and regulatory purposes. This work will be further discussed and developed within the OECD expert group on non-genotoxic carcinogenicity IATA development.

Keywords: carcinogenicity testing, uncertainty analysis, defined approaches, IATA

1 Introduction

1.1 History of the rodent cancer bioassay

Experimental carcinogenesis was initiated in the first part of the 20th century, when Yamagiwa and Ichikawa first developed an experimental model to study the pathogenesis of carcinoma and then used it to confirm the causal effect between exposure and cancer (Yamagiwa and Ichikawa, 1918). The experiment was performed by applying coal tar onto a rabbit's ear in order to confirm Sir Percival Pott's observation, published in 1778, on the link between the exposure to soot and the increased incidence of scrotum squamous cell carcinoma in chimney sweepers. The results from this experiment were confirmed in 1933, just after the identification of polycyclic aromatic hydrocarbons, by using benzo(a)pyrene to induce skin papillomas in mice (Cook et al., 1933).

These early experimental studies were undeniably a milestone in the comprehension of the carcinogenesis process. They also marked the beginning of the use of toxicology to verify epidemiological evidence from occupational exposures. The first carcinogenesis animal assays were set up to resemble occupational exposure, using skin cancer as the endpoint. The skin carcinogenesis model was also used to replicate the multistep carcinogenesis process in a two stages initiation-promotion experimental study, which was published in 1941 (Berenblum, 1941).

In the following years, in the absence of specific guidelines, animal studies were performed in a limited number of animals,

Disclaimer: The scientific views presented in this paper are those of the authors alone and do not necessarily reflect official views of their respective institutions.

Received August 28, 2016;
Accepted October 1, 2016;
Epub October 24, 2016;
https://doi.org/10.14573/altex.1608281

exposed for a limited period to a limited number of doses that were sufficiently high to observe the adverse effect, thus confirming the epidemiological findings.

The 2-year rodent carcinogenicity bioassay (RCB) was proposed for the first time in 1949 by the Food and Drug Administration (FDA) and a guidance was provided in response to the need to standardize the experimental protocol due to the high number of chemicals to be tested (Jacobs and Hatfield, 2013). The guidance was specifically issued for industry to assess the toxicity of chemicals in food. It was based on a long-term feeding protocol and two species were required, i.e., albino rats, treated for a period of two years, and a non-rodent species, dog or monkey, treated for a period of one year. Over the years, oral administration became the main route of animal exposure. The first studies by inhalation were performed in 1958, but since then, due to the high costs and complexity, this route of exposure has been considered only in a small percentage of animal bioassays.

However, the standardization of the testing approaches was not put into place until the end of the 1970s. As an example, several long-term cancer experiments were performed between 1960 and 1970 to test pharmaceuticals in dogs and monkeys with treatment schedules extending up to 7 or 10 years (Alden et al., 2011). All the information from these non-standard animal tests concurred to provide the evidence for carcinogenesis classification in the International Agency for Research on Cancer (IARC) evaluation process, which had started in 1971. It was in response to the lack of quality and scientific integrity of several toxicology studies, as revealed by FDA's investigations into laboratories of animal research, that the FDA decided, at the end of the 1970s, to adopt federal regulations for conducting non-clinical studies of chemicals. This decision prompted the development of international guidelines and the implementation of Good Laboratory Practices (Baldeshwiler, 2003).

The current 2-year carcinogenicity bioassay has been adapted from the original FDA protocol, with rodents as the preferred species, an expansion of the number of animals required at each dose level, and at least three treatment doses. Whilst the test protocol has undergone some refinement over the years, the current experimental protocol still reflects the original experimental design developed to resemble occupational exposure (though with imperfect concordance to human development and full life exposure, Fig. 1 and Tab. 2) and to highlight the carcinogenic activity of genotoxic chemicals.

Indeed, the fields of genotoxicity and mutagenicity testing developed concurrently with the 2-year RCB. Several known *in vivo* carcinogens were unsuccessfully tested in the *in vitro* mutagenicity assays, since most of the tested chemicals required metabolic activation, which was not supported by the early mutagenicity models (Mahadevan et al., 2011). It was only at the end of the 1960s, with the metabolic improvement of the AMES assay, that the correlation between mutagenicity *in vitro* and carcinogenicity *in vivo* developed further to become the basis of the current testing strategy for the identification of carcinogens.

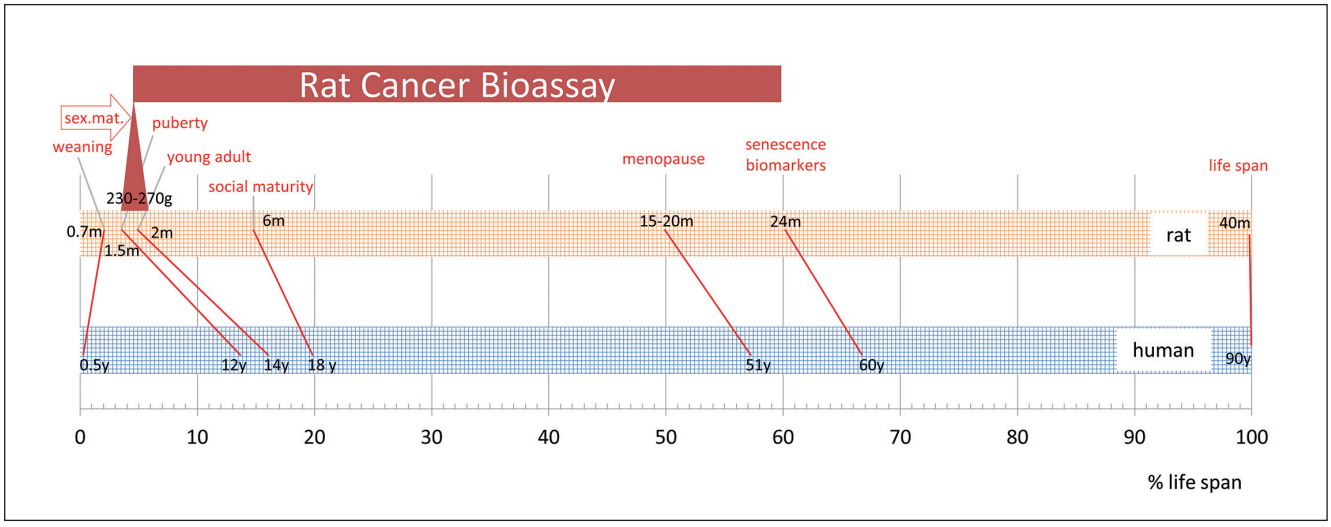

Fig. 1: Imperfect concordance of rodent 2-year cancer bioassay with human development and full life exposure

Compared to humans, rodents have a shorter and accelerated early life. According to OECD TG 451, test animal (rodents) treatment should begin as soon as possible after the weaning period and before 56 days of age and the weight variation for each gender should be minimal, not exceeding ±20% of the mean weight of all the animals within the study. Test animals are often selected by weight of 230-300 g as an indicator of age. However, e.g., male rats weighing 230-270 g may differ in age between day 49 (periadolescent) and day 70 (young adult). Experimental groups may therefore include animals at different stages of early development. The corresponding human age ranges between 12.5 to 14 years (still juvenile, not adult). Female rodent reproductive senescence (menopause) and overall senescence (senescence biomarkers) is reached earlier in rats compared to humans. In summary, the treatment starts when some of the animals are still in the periadolescent period and continues when animals have reached (and may be past) reproductive senescence (Dutta and Sengupta, 2016, Sengupta 2013, Demetrius 2006). Not covered is: prenatal exposure, postnatal exposure and development until sexual maturity and, finally, later senescence [g, grams; d, days; m, months; y, years; sex.mat.= sexual maturation].

At present, for current chemical risk assessment for human health purposes, the RCB constitutes the *in vivo* "reference" data and is considered to be the current "gold standard" for the identification of carcinogens. It is both this assay and this assumption that require closer retrospective examination and uncertainty assessment to facilitate the development of more appropriate and relevant fit-for-purpose carcinogenicity testing strategies in the 21st century.

1.2 Why do we need to describe the complexity and other uncertainties of the current *in vivo* carcinogenicity testing and assessment approach?

In principle there are three reasons to engage with this discussion:

1) Foster the interest in improved and newly defined in silico and in vitro approaches

It is already broadly recognized that besides ethical and consequent policy concerns with animal testing, there are many practical regulatory needs for 3R methods (for instance NAS, 2007; OECD, 2005; Bal-Price et al., 2015) and many major international scientific conferences and advisory bodies specifically support this aim, including the development of alternatives to the 2-year RCB (e.g., 9th World Congress on Alternatives and Animal Use in the Life Sciences[1]; ECHA, 2016; EUROTOX, 2015[2]; ESTIV, 2016[3]; EUSAAT, 2016[4]; COC, 2016). One of the most important goals is to increase the testing throughput in order to improve

- the availability of regulatory test data for many chemicals,
- the effectiveness of substitution of the more hazardous chemicals by providing reliable data for an ample set of potential alternative chemicals, also in the low tonnage production range or even for green chemical engineering (Maertens et al., 2014),
- the assessment of mixture toxicity,
- the assessment of environmental media including the use of bio-analytics to complement chemical analytics (Schroeder et al., 2016),
- approaches to cross-species extrapolation for ample coverage of environmental toxicity (Groh et al., 2015),
- the assessment of the multitude of nanomaterial compositions, forms and size distributions,
- the possibility to retest chemicals according to progress in the development of scientific and toxicological understanding.

In addition to these practical needs, a critical mass of concern regarding the scientific uncertainties of animal test results and their regulatory utility also has been steadily accumulating in recent years (e.g., Basketter et al., 2012; Paparella et al., 2013; Hartung, 2013; Leist et al., 2014; NAS, 2015). Amongst these publications, a poor reproducibility of 57% for the 2-year RCB (Gottmann et al., 2001), and for pharmaceuticals in humans, high false negative rates and high false positive rates are reported (e.g., Alden et al. 2011). Yet, it is the specific, detailed and scientifically more rigorous analysis of the deep uncertainty and complexity inherent to animal testing and assessment that may most substantially contribute to stimulating interest in new approaches.

2) Define what type of effects we want to predict with the new defined in silico and in vitro approaches and select the respective reference chemicals and data

Carcinogenicity is a relatively broad term and a common understanding is needed as to what constitutes a relevant carcinogenicity finding. Various data, i.e., human, animal, *in vitro*, mechanistic data, and various possible ways to assess and integrate these, need to be considered (Annys et al., 2014). Not all approaches may be of equal merit for the final goal of risk management.

3) Decide on acceptance criteria for the new defined in silico and in vitro approaches

The correlation of defined *in silico* and *in vitro* results from approaches with standard reference results is limited by the complexity and other uncertainties of standard RCB reference results. A clear and harmonized picture is needed on these complexities and uncertainties to define a benchmark for the required performance of new, defined *in vitro* and *in silico* approaches.

This article starts from a conceptual discussion of point 3, which is followed up with a proposal for a systematic review of complexities and uncertainties with regard to carcinogenicity. It finally proceeds to a long term view on regulation based on defined *in silico* and *in vitro* approaches.

2 How to decide on acceptance criteria for new, defined *in silico* and *in vitro* approaches?

Usually, validation of new *in vitro* methods starts from a clear test definition including the exact scientific purpose, SOPs and prediction models and estimated reliability, i.e., reproducibility, of the new *in vitro* method and the relevance, i.e., correlation to "gold standard" animal test reference data. As far as available also human reference data and mechanistic information is used to support validation (OECD, 2005; Hartung et al., 2004). With advancing *in vitro* toxicology that targets increasingly complex endpoints, like earlier skin sensitization integrated approaches to testing and assessment (IATA) (OECD, 2012a) or most recently non-genotoxic carcinogenicity (Jacobs et al., 2016), it is not just individual methods that need validation, but an optimum combination of several *in silico* and *in vitro* methods. In the most recent OECD terminology, such a combination of methods that includes a prediction model integrating various *in silico* and/or *in vitro* method read-outs is called a defined approach (OECD,

[1] http://www.wc9prague.org/
[2] http://eurotox2015.com/home/83/
[3] http://estiv.org/
[4] http://eusaat-congress.eu/

2016). In an earlier OECD *in silico* methods document a similar concept was described as the evolvement of quantitative structure activity relationship (QSAR) models, relying just on chemical structure input data, to quantitative activity activity relationship (QAAR) models, including structural and biological input data (OECD, 2014). It could be understood that a defined approach represents also what earlier has been called an integrated testing strategy (ITS), but includes also non-testing information. The understanding is that defined approaches are building blocks within IATAs (OECD, 2015, 2016). An IATA is a conceptual guidance that explains under which circumstances how and why one or other of the defined approaches should be applied and integrated with further information, such as exposure, for regulatory decision-making purposes. So, the largest unit that conceptually may be validated is the defined approach. It remains a challenge to validate defined approaches due to the multitude of possibilities to integrate read-outs from several methods, from Boolean logic to probabilistic and fuzzy logic (Hartung et al., 2013; Jaworska et al., 2011). However, an IATA is perhaps too complex to be validated in the strict sense (Hartung et al., 2004; OECD, 2005) and the terminology "evaluation of fitness" appears more appropriate for IATAs, as it should allow a semi-quantitative or qualitative evaluation.

In any case – be it for the validation of methods or defined approaches, or for the evaluation of fitness of IATAs – in principle three sources of reference data may be engaged, i.e., animal, human, and "mechanistic" data. To understand the relative weighting to be allocated for each of the three sources of reference data (in the context of assessing the acceptability of the new method, new approach, or IATA), a systematic analysis of the uncertainties for the information contained within each of these three classes is required. Since the ultimate goal is *not* to predict whatever reference data and protect rodent health, but is to optimize the tool kit we have to better protect human health and the environment, we need to go beyond the current reference data and examine the bigger picture. Within this broad overview, we need to systematically assess 1) the reliability of reference data and 2) the correlation of these reference data to the target of evaluation, be it human health or the environment.

In principle, three criteria decide on the acceptability of the new approach (Fig. 2):

A) The results of the new approach should be at least as reliable as the reference data.

B) The correlation of the new approach's results with the reference data should be close to but not better than the reliability of the reference data.

These points are self-explanatory and well developed in *in silico* toxicology. For example, point B is an issue usually carefully considered for QSAR development: "over-fitted" models, i.e., models with a standard error of estimate smaller than the experimental error of the biological data have to be avoided, otherwise the performance estimate for the new defined approach would not be robust, as it would change with each new data-pair included in the analysis (OECD, 2007).

However, some conceptual clarification is useful to discuss aspects around uncertainty (see Tab. 1). So far, in the hazard and risk assessment context, the focus was on the distinction of pure uncertainties, which may be reduced with further knowledge versus variability of results that cannot be reduced with further knowledge (WHO, 2014; ECHA, 2012). In the context of risk governance, the term "complexity" is used to describe effects triggered by many causes, inter-related like a network – and "ambiguity" is used for uncertainty stemming from the plurality of scientifically legitimate viewpoints (Renn et al., 2011; IRGC, 2005). Finally, "ignorance" addresses the fact that our toxicological knowledge is continuously evolving – so while we can fill some data gaps in our understanding, new areas of uncertainty will emerge. Thus our "ignorance" also changes, but does not disappear (EEA, 2010).

From this perspective, for the assessment of reliability of animal testing reference data with regard to the points A and B, two aspects need consideration: 1) variability due to protocol variants and 2) variability due to limited reproducibility of strictly defined and identical protocol variants. Estimates from identical

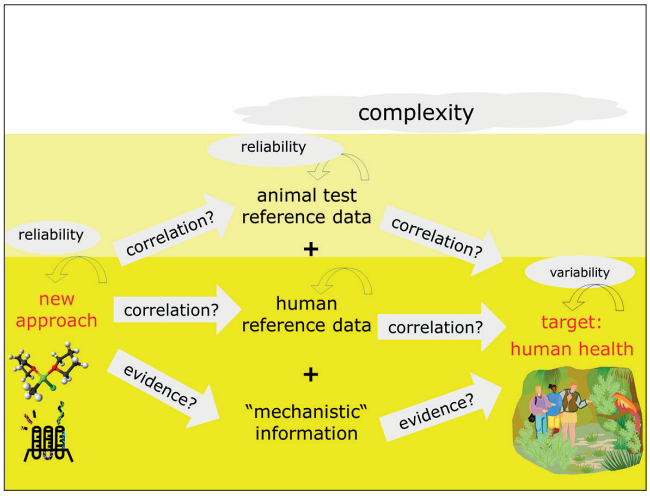

Fig. 2: Validation or evaluation of the fitness of new approaches to testing and assessment

To decide on the acceptability of new approaches, three conceptual criteria shall be engaged: A) The new approach's results should be at least as reliable as the reference data; B) The correlation of the new approach's data with the reference data should be close to but not better than the reliability of the reference data (to avoid "overfitting"); C) the higher the reliability of the reference data and the higher the correlation of the reference data to the target of evaluation, the more weight should be given to these reference data, which would then be better described as line of evidence. Complexities and ambiguity around respective reliability and correlation estimates must not be underestimated and shall be taken into account as a further important aspect of intrinsic uncertainty and shall inform the definition of acceptance criteria for the new approach. It is expected than in many cases all three lines of evidence have to be integrated, but that evidence based on mechanistic information (AOPs) and very well characterized human reference data, even if only little is available, will often be most important (most "golden") for assessing the validity or fitness of new approaches.

Tab. 1: Terminology of uncertainties

Term	Explanation	Example	Reference
variability	uncertainty intrinsic to biology, cannot be reduced with further knowledge	different species and protocol variants will produce results that vary from each other to a certain extent, may be described as probability distribution	WHO, 2014; ECHA, 2012
"pure" uncertainty	uncertainty due to limited knowledge, can be reduced with further knowledge (at least theoretically)	knowledge of variability of results from different species and protocol variants is uncertain , may be described qualitatively and quantitatively in terms of a confidence interval to the probability descriptor	
reliability	variability plus uncertainty	see above	OECD, 2005
complexity	uncertainty stemming from multi-causal effect relationships	derivation of a reference value is the result of a series of decisions in the process of data generation, assessment and integration, e.g., decisions on animal numbers and top doses, definition and grouping of tumors, statistics and use of historical controls	Renn et al., 2011; IRGC, 2005
ambiguity	uncertainty stemming from the plurality of scientifically legitimate viewpoints	different expert groups weighting and integrating the same data differently and thereby come to different conclusions	
ignorance	any uncertainty that we are not aware of and cannot name: 'what we don't know that we don't know'	epigenetic modes of action were "ignorance" until relatively recently, but are now moved to being recognized as uncertainty	EEA, 2010

protocol variants are difficult to obtain since replications of fully identical protocol variants are rarely available. Biological variability, in terms of species and strains, contributes to variability in toxicity tests just as caging, handling, feed, exposure regimes and routes, and others. Therefore, knowledge of variability of results from specific protocol variants is uncertain. It should be considered that in principle it is not easy to decide if a specific protocol variant of the standard animal test is more relevant than another, especially *a priori* to testing. Furthermore, many protocol variants are in principle covered by the frame of the OECD Carcinogenicity Test Guideline (TG). Therefore, from a regulatory point of view, all the protocol variants that are within the frame of the TG are in principle acceptable. Moreover, from a scientific point of view, these protocol variants may be considered to represent some range of natural variability. Therefore, variability due to these protocol variants may be considered as TG inherent and a useful and practically reasonable estimate for reliability of animal test reference data.

Also, uncertainty stemming from the various possibilities to generate, assess and integrate data should not be underestimated. Any classification or reference value is the result of a series of decisions in the process of data generation, assessment, and integration, e.g., decisions on animal numbers and top doses, definition of tumors, grouping of tumors, use of statistics and historical controls. This means that any result of an assessment is multi-causal or, in other words, there is "complexity" underneath each assessment result. Furthermore, there is often more than one scientifically defensible way to assess and integrate data, different expert groups may have different opinions, often arising from slightly different remits, which is normal, natural and important for the evolution of science and scientific debate, but a challenge for regulatory decision-making, where a clear decision one way or the other is required. The latter was termed "ambiguity".

It is acknowledged that uncertainty and variability needs to be scientifically assessed and managed with a view to precaution, likewise complexity needs to be assessed scientifically and managed with a view to optimize robustness. Ambiguity (e.g., around uncertainty, variability, and complexity) may be resolved by a stakeholder discourse-based strategy (IRGC, 2005), which is a consensus-finding task, and there are several multivariate science policy tools, such as MCDA (multi-criteria decision analysis) that have been developed to capture, elucidate, and identify uncertainties, transparently facilitating the democratization of decisions at the science-policy interface (e.g., Linkov et al., 2013; Sailaukhanuly et al., 2013). The uncertainties concept of "ignorance" reminds us that science is constantly evolving, and underpins the more cautious language often employed by scientists. However, in order to support consensus-finding for complex issues, it is important to note that science may never provide definitive answers, but answers in terms of probabilities (like weather forecasts). Therefore, it is finally a science-informed (and not a science-based) policy decision that is necessary to advance regulatory toxicology.

A systematic approach to start characterizing reliability, complexity and ambiguity of the animal reference data and assessments is provided in Section 3. For human reference data, quality assessment criteria are specific to different disciplines, i.e., clinical, epidemiological, etc. Therefore, the necessary reliability estimate of such data needs a careful case-specific

approach. ICH guidelines can provide a frame for such an analysis of clinical data. Human variability, potential for selection-bias for group composition, information bias for retrieval of information and confounding by hidden factors are generally recognized key terms for the analysis of epidemiological data and studies (see e.g., Blettner et al., 2001), and the Bradford Hill criteria are often used to assess the quality of the epidemiological evidence: strength of association, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, analogy (Hill, 1965). Adami and colleagues have made an attempt to provide a framework for the integration of epidemiological data with toxicity data (Adami et al., 2011), and Samuel and colleagues provide a review of guidance documents for the assessment of methodological quality of human observational studies (Samuel et al., 2016). The review may provide an entry point for identifying the most appropriate case-specific approach. A relevant case study example of the literature screening and integration of epidemiological and toxicological information for cancer outcomes of concern for three pesticides was recently conducted by the WHO FAO Joint Meeting on Pesticide Residues[5]. Also, literature on human variability may be critical for conclusions (Zeise et al., 2013; WHO, 2014; McNally et al., 2015).

In any case, the correlation between results from new approaches and reference data will be influenced by two aspects, i.e., the reliability (i.e., variability and uncertainty) of the new approach and of the reference data, and the correlation of mean values from the new approach and the reference data. Appropriate statistical concepts supporting analysis according to point B need to be described. However, also more pragmatic qualitative analysis approaches could eventually be adequate. Especially, for characterizing how strong the evidence is that the new approach provides the desired mechanistic information, a scientific Weight of Evidence analysis, which will also be very case specific, is necessary.

C) The higher the reliability of the reference data and the higher the correlation of the reference data to the target of evaluation, the more weight we should give to these reference data, which would then be better described as line of evidence.

Evaluating point C means to integrate knowledge on reliability of reference data and their relevance for the target of evaluation and conclude on the relative weight the individual lines of evidence should have for assessing the fitness of the new approach. Further consideration is necessary on how to assess the relevance of reference data and information for the target of evaluation. A systematic approach to start characterizing reliability, complexities, ambiguities and relevance of standard reference animal test data is provided in the next section. For characterizing the evidence that defined mechanistic data may lead to an adverse human health outcome, the current adverse outcome pathway (AOP) concepts and guidance shall be engaged. Complexities and ambiguities around respective reliability and relevance estimates must not be underestimated. There are always several ways to assess and integrate complex data and it may not be easy to objectively give preference to the one or other approach. This shall be duly taken into account as important aspects of intrinsic uncertainty, and shall inform the definition of acceptance criteria for the new approach. It is expected than in many cases all three lines of evidence (animal data, human data, mechanistic information) need to be integrated for the assessment of fitness of new approaches, but that evidence based on mechanistic information (modes of action (MoAs) and AOPs) and very well characterized human reference data, even if only little is available, will often be most important (most "golden").

3 What are the uncertainties affecting our recognition of carcinogenicity and respective effect levels today?

The analysis of these uncertainties can be structured along the OECD guidance document on reporting of defined approaches within IATAs (OECD, 2016) (Tab. 2). Conclusions like "sufficient evidence of carcinogenicity in animals" need integration of complex data sets and can be described as data interpretation procedures similar to those expected from new *in silico* and *in vitro* approaches. Such a structured characterization may support appropriate reference data selection and a comparative evaluation of uncertainties with potential new defined *in silico* and *in vitro* approaches. On this basis – as outlined in Section 2 – a decision on the acceptability of new approaches can be taken. The uncertainties and complexities around the current definition of "sufficient evidence of carcinogenicity in animals, including point of departure for risk assessment" were specifically analyzed for each of the four elements defining a testing and assessment approach: 1) the endpoint addressed, 2) the rationale underlying the construction and interpretation of the approach, 3) the description of the individual information sources constituting the defined approach, 4) the data interpretation procedure applied:

1) The endpoint assessed, e.g., how to define the maximal tolerated dose (MTD) and lung overload? Are we interested in effects above the MTD? Which MoA are of interest, e.g., local irritation leading to carcinogenicity with or without genotoxicity (e.g., formaldehyde)? Is carcinogenicity of (very) persistent and (very) bioaccumulating substances of interest?
2) The rationale underlying the construction/interpretation of the approach, e.g., what are the uncertainties related to the difference of rodent to human life span and life stages (Fig. 1), and related to the high to low dose extrapolation? Are we only interested in multi-species, -strain, -study, -sex, -tissue neoplasms, or is this criterion not adequately protective for humans, since non-genotoxic mechanisms also should be included? How does the multitude of potential protocol

[5] See Fig. 1 in http://www.who.int/foodsafety/jmprsummary2016.pdf?ua=1 and forthcoming WHO monographs on glyphosate, malathion and diazinon

Tab. 2: Application of the *OECD guidance for reporting defined approaches within IATAs (OECD, 2016)* for systematic description of uncertainty and complexities of animal based carcinogenicity evaluation

Defined Approach	**Known complexity and other uncertainties associated with the application of the approach:** uncertainty from approach structure, information sources and benchmark data (shaded in light grey)
GENERAL INFORMATION	
Identifier: sufficient evidence of carcinogenicity in animals including PoD	
Reference to main scientific papers: GHS; ECHA CLP Guidance	
ENDPOINT ADDRESSED	
Endpoint: induce cancer increase incidence and/or malignancy reduce time to tumour benign + malign including dose-response	What is the maximum tolerated dose (MTD), may different definitions be defensible and are we interested in effects above the MTD? How to define local tissue MTDs (i.e., lung overload)? What MoA do we want to diagnose, e.g., neoplasms resulting from chronic inflammation due to local irritation effects in the lung or local genotoxicity? Do we want to diagnose (very) persistent and (very) bioaccumulative substances for carcinogenic potential? They are in any case already of high concern, they may not be able to reach steady state, some may be poorly water-soluble and therefore problematic for *in vitro* testing. The site of neoplasm in the lab animal is not necessarily predictive for the site of neoplasm in humans. In which situations does this matter, and in which does it not?
Species: human predictivity is the goal	
Additional information: genotoxic and non-genotoxic MoA	
DEFINITION OF THE PURPOSE AND REGULATORY RELEVANCE	
Hazard assessment/characterization for classification and potency estimates Definition of PoD for estimating acceptable exposure level (for risk assessment)	
RATIONALE UNDERLYING THE CONSTRUCTION OF THE DEFINED APPROACH (i.e., assumptions that the RCB based approach is valid model for human cancer hazard assessment)	
– Most human carcinogens are also carcinogenic in standard rat and mouse carcinogenicity studies	What do we know about the sensitivity of the RCB to predict human carcinogens? Depending on the need for positive data in 1 or 2 species for defining a carcinogen as positive in animals, a sensitivity of 50-90% was reported for 10 IARC confirmed human carcinogens for which adequate standard animal test data were available (Ennever and Lave, 2003). Are these data reliable? What is the false negative rate, specifically for pharmaceuticals (Alden et al. 2011)? In any case, also false positives are of concern (see comment on predictive capacity below). Pre- and early post-natal exposure is likely to be critical for carcinogenesis, therefore it is unlikely that the standard RCB (not covering this period) is fully predictive for real life human carcinogenesis (review conclusion in Downes and Foster, 2015).
– Higher dosing	By default, doses of up to 1000 mg/kg bw day should be applied in the RCB in order to allow the observation of significant carcinogenesis using not more than 50 male and 50 female animals/dose group. How relevant is this dosing regimen for the assessment of potential carcinogenesis in real life human exposure situations, including the general public and workers?
– … over young to adult life span	RCB protocols were originally designed to address occupational exposure; however, the developmental phases of rodents and humans are poorly concordant (see Fig. 1). There are critical developmental windows of exposure and test animals are likely to be at different stages of development at the initiation of the test (particularly when bought in). Reproductive senescence and overall senescence is much faster in rodents compared to humans. Thus, how can we adequately utilize this assay for assessing the carcinogenic properties of potential endocrine disruptors?
– … which is shorter than the human life span	Is the development of neoplasms in the rodent life span a sufficiently sensitive endpoint to identify a potential hazard in the aged human population?
– Neoplasms are more likely to be relevant, if identified in more than one species,	Effects seen in single species are more likely to be pathway specific. However, these may nevertheless be relevant for humans (Downes and Foster, 2015).

– At multiple sites, in more than one study, in both sexes	Are genotoxic chemicals more likely to show multiple site carcinogenicity? How much do we lose of the value added of the carcinogenicity study (compared to genotoxicity tests) if we require multiple strains, species, multiple studies, multiple site neoplasms, rare neoplasms seen for species being tested? (And when developing new *in vitro* NGTxC IATAs – should we focus on reference chemicals showing these multiple effect characteristics? (Gray et al., 1995)).
– The dose causing a significant increase of neoplasm in the animal study is a relevant PoD for deriving acceptable human reference doses.	Human neoplasms often have multi-factorial causes, i.e., a combination of culture, diet, life style, genetic background, and co-exposure, but RCBs test for effects from single substance exposure. RCBs are not designed to deliver information on normal variance of effects and effect level within a human population and human variance can be very high and at population level no thresholds exist for any type of effect (WHO, 2014; Schneider et al., 2005) The low variance within an animal test does not mirror the variance between studies and the multitude of possibly relevant study designs, animal strains and species. What is the decision basis for the selection of an appropriate test design? What is the correct PoD for limit value derivation? What are the uncertainties related to exposure route extrapolation? These may relate to kinetic and metabolic differences, e.g., first pass effect, local effects seen with gavage, absorption and metabolism differences between dietary and gavage exposure, stress from gavage application (Vandenberg et al., 2014; Proctor et al., 2007). The presence of inconsistent study results is the rule in many assessments (e.g., 20% of animal studies appear inconsistent with regard to dose-response relations for the endpoints body weight, liver weight, kidney weight, erythrocyte count (Paparella et al., 2013 and references cited therein)). The standardization of TGs and assessment rules is limited. The high amount of data and high complexity of data integration and interpretation and natural "biodiversity" of human scientists easily leads to different expert groups coming to different conclusions (e.g., Schneider et al., 2009: 12 experts + 11 *in vivo* studies: for 2 studies same Klimisch score (KlimS) categories by all experts, for 8 studies ratings over two neighboring KlimS, for 1 study ratings over all three KlimS; Rudén, 2001: 29 assessments for trichlorethylene with conclusions distributed over 4 categories from clear negative to clear positive carcinogen; CRD, 2013: different ADI derivation by EFSA and JMPR for 23/ 57 substances).

DESCRIPTION OF THE INDIVIDUAL INFORMATION SOURCES USED	
Mechanistic basis including coverage of the AOP: Default assumption: All human relevant MoA/AOPs are covered in the animal tests	How many human relevant pathways cannot be observed in the RCB? For example, although animal models have been created for drug therapeutic purposes (Young et al., 2009), the RCB is not a suitable model for non-Hodgkin's lymphoma in humans. Also, thyroid mechanisms are species-specific. Rats do not have gall bladders. How many rodent pathways are not relevant in humans? Such as peroxisome proliferation leading to liver tumours, forestomach and thyroid cancers (Boobis et al., 2006; Proctor et al., 2007; Meek et al., 2014). Age associated tumors are common and variable amongst rodent species and strains, e.g., Leydig cell tumors (Creasy et al., 2012; WHO, 2015). Usually very limited information on MoA/AOP from standard animal studies may lead to false conclusions for human hazard assessment (for instance atrazine, DEHP etc., discussed in Jacobs et al., 2016).

Description:
Sufficient evidence of carcinogenicity in animals:
Neoplasms in standard carcinogenicity TG

– in two or more species, or
– two or more independent studies, or
– in one study carried out at different times or in different laboratories or under different protocols, or
– both sexes in single species in one well conducted study, ideally under GLP, or
– single species, one sex when malignant neoplasms occur to an unusual degree with regard to incidence, site, type of tumour or age at onset, or when there are strong findings of tumours at multiple sites,
– overall NOAEL/LOAEL or BMD estimate for effect.

Response measured: Within each study: Dose-response – for each organ: number of animals with neoplasm (benign + malignant) – total number of animals with neoplasm (benign + malignant) – latency – correlation with other potentially related toxicological findings (e.g., organ weight changes, clinical biochemistry or haematology findings)	Histological nomenclature and standardized diagnostic criteria for neoplasms may vary, e.g. over time. There is a potential for subjectivity in histological endpoint analysis (i.e., an image); pathology working groups may be necessary for arbitration in case of disagreements in pathology review. Uncertainties may be due to a diagnostic drift, i.e., increased awareness of a lesion by the pathologist leading to gradual change in nomenclature or severity grading. This is especially an issue with large studies requiring evaluation over a prolonged period of time. There is potential for bias, e.g., a histopathologist knowing gross pathology and organ weight effects in each animal can improve sensitivity of his analysis at the cost of potential bias; furthermore, differences of observer, equipment, timing of investigation may introduce bias; this may be minimized with randomization. There are limitations for the blinding of the analysis: e.g., analysis of baseline histopathology in the control group is often necessary and usually the analysis is started with control and high dose group (OECD, 2012b).
Prediction model: Statistically significant increase of organ-specific neoplasm compared to control by – pair-wise comparison, or – trend test, or – size of effect and confidence interval at specific doses (BMD approach) AND	Which neoplasms (usually including benign & malign, but excluding metastases) are usually combined for analysis and could different approaches also be scientifically defensible? Which statistical approaches are applied and how many others could also be relevant and defensible: Non-parametric tests (Chi-squared, Mann-Whitney, Cochrane Armitage), one tailed vs two tailed, p-value, compensation for multiple testing analysis, also over several studies (multivariate data analysis and meta-analysis)? How far can randomisation and independence of animals and analysis be granted? How does the number of animals affect RCB results? May increasing animal numbers from 50 to 200/dose group result in statistically significant dose-response trends for nearly all RCBs (Gaylor, 2005)? How does the number of histological slices/organ and type of cut (transversal/ cross sectional) influence the probability to identify tumours? Do non-monotonic dose-response (NMDR) relationships exist and would they be identified by the RCB? Reproducible NMDRs may be observed with non-monotonic kinetics, e.g., if uptake is reduced due to agglomeration at high concentrations; or if at higher concentrations high cytotoxicity or general toxicity covers the more specific effects; or where more than one mechanism operates at differing dose levels resulting in the same mode of action of the endpoint being affected due to different mechanisms within that MoA; or due to experimental variability in the low dose range. The detection of a true NMDR relies upon appropriate study design (with sufficient dosing in low dose range; just 3 doses with MTD and 50% and 25% of MTD is not sufficient; Borak and Sirianni, 2005), and the sensitivity of the method (depending on sample size and background variability) and technical laboratory expertise. Accepting NMDRs as "true observations" depends on our understanding of what constitutes an adverse effect (see Section 4 of this manuscript). NMDRs were observed as organ specific (e.g., Cadmium chloride), sex specific (e.g., dioxin) and may also be due to experimental variability (Borak and Sirianni, 2005). If NMDRs arise they "*may affect the appropriateness and ability of conventional testing to identify where the overall experimental threshold lies*" (EC, 2013; see also U.S. EPA, 2013; Testai, 2015; Andersson, 2015).
outside of historical control data range AND	Comparison with historical control range is often the reason for disregarding observed neoplasms. However critical details are often difficult to assess and therefore rarely addressed in evaluations: Changing trends in historical control data may be due to genetic drift, caging protocols, diet, study duration, survival differences, etc. Study conditions and analysis techniques may change.
identified in two or more species, or two or more independent studies, or both sexes in well conducted study or single sex with unusual degree AND	Rats are the standard test model used in toxicology; it is reported that classification and limit value derivation for pesticides are rarely significantly influenced by results from RCBs with mice (Billington et al., 2010) and for pharmaceuticals, neoplasm findings in mice alone did not lead to regulatory action (Van Oosterhout et al., 1997)? Are we interested in multiple species/study/site etc. effects? (see above rationale for construction)
no MoA contradicting human relevance AND	What is sufficient evidence for excluding human relevance of MoA? (Boobis et al., 2006; Meek et al., 2014)
acceptable data quality assurance systems applied: GLP, independent review	Are all critical aspects covered by QA? Contaminant residues/impurities in laboratory rodent diets and test substance and bedding may affect the outcome of a RCB.

Metabolic competence: competent	Knowledge of human to laboratory animal differences in metabolic competence is usually limited (exception pharmaceuticals). Gender differences in chemical metabolism may be expected (Lewis, 2002).
Status of development, standardisation, validation: standardised via OECD TG; no international validation in terms of reproducibility and human relevance available	
Technical limitations and limitations with regard to applicability: Very low testing and assessment throughput: at least 400 animals, €800.000, 2 years testing + 2 years analysis. Mixture testing problematic for pragmatic reasons and, since mixture may not achieve MTD, the sensitivity of the method may not be high enough. Nanomaterial testing problematic for pragmatic reasons, due to potential multitude of nano-forms and -distributions.	Standard RCB based assessment of single substance toxicity may be of limited relevance for human exposure situation to products and environmental contaminants. Nanomaterial toxicity may depend on size distribution, form, impurities, and other – and is of high policy concern and this high number of variations cannot be addressed by RCB. For nanomaterial toxicity, dosimetry and aggregation are potential issues of uncertainty.
Reliability (within and between laboratories): No internationally agreed validation available	What is the concordance between replicate assessments? A value of ~ 57% was published by Gottmann et al. (2001) based on the carcinogenic potency database, which contains two components: the National Cancer Institute / National Toxicology Program (NCI/NTP) database and the literature database. For 121 chemicals carcinogenicity studies were available in both components; for each of these chemicals, one assessment was carried out based on the studies in the NCI/NTP part and another assessment based on the studies in the literature database; a substance was considered positive if a positive result was obtained in at least one experiment.
Predictive capacity: No internationally agreed validation available	What is the false-negative rate and false-positive rate for pharmaceuticals? How reliable is the evaluation drawn from the physicians' desktop references database, specifically assessing consistency between the section indicating enhanced human cancer risk and the section on animal testing result (Alden et al., 2011)?
Proprietary aspects: no	
Proposed regulatory use: risk assessment and classification of chemicals	
Potential role within an IATA: To be integrated with: 1) human epidemiology, human clinical data 2) genotoxicity tests, 3) MoA tests	

DATA INTERPRETATION PROCEDURE APPLIED	
Expert based WoE evaluation; guidance documents, e.g., OECD, ECHA, IARC; templates and guidance for MoA relevance analysis available: http://www.who.int/ipcs/methods/harmonization/areas/cancer/en/	
Does the prediction include an assessment of uncertainty? Conclusion is usually deterministic, i.e., yes/no	The usual deterministic conclusion (i.e., carcinogen or non-carcinogen, based on sufficient or limited or inadequate evidence) does not reflect scientific reality, the latter is rather probabilistic.
Sometimes available: – transparent WoE assessment – MoA analysis – explanation for use of assessment factors for acceptable exposure level derivation	Data are highly complex and experts are "biodiverse" by nature (see also last paragraph in right column to "Rationale underlying the construction of the approach/for interpretation of predictions")
Feasible: – probabilistic hazard assessment to define target human dose (HD_M^I) at which with x% confidence I% of the population will have an increase of risk for neoplasms by M% (WHO 2014, APROBA spreadsheet)	The uncertainty, i.e., ratio of P95/P5, for HD_M^I is increasing the lower I (the accepted residual percentage of population under risk), spanning several orders of magnitude at level 10^{-6}. Not quantifiable uncertainties need additional characterization. The database used for the development of probabilistic assessment factors also contains qualitative and quantitative uncertainties.

designs (all within the frame of the OECD TG), including various species, strains, exposure routes and more, affect variability of the results and selection of points of departure? What is the proportion of human variability relative to all experimental variability?

3) The description of the individual information sources constituting the defined approach, e.g., what are the uncertainties around the nomenclature of neoplasms and the related diagnostic criteria, subjectivity of histological assessment, bias and blinding issues? How does the number of animals and use of various statistical methods and the use of historical control data and trends affect the outcome? What is the chance to identify relevant non-monotonic dose-response relationships? What is sufficient evidence to exclude human relevance? Do quality assurance systems cover all critical aspects, e.g., including lab rodent diet? What do we know about concordance between replicate studies, "false" negative and "false" positive rates?
4) The data interpretation procedure applied, e.g., what is the influence of limited standardization and complexity of data integration on the variability of results? In how far is the deterministic conclusion (yes/no) for carcinogenicity adequate for regulatory purpose? How uncertain is the final human reference dose?

Table 2 lists and references all these aspects and it can be seen that, although preliminary, there is utility in using such a template for initiating a systematic analysis, discussion and further development within the OECD expert group on non-genotoxic-carcinogenicity IATA development. It appears to be suitable to support, structure and substantiate the discussion on the complexity, i.e., the potential multitude of approaches for integrating and interpreting data from standard animal testing approaches. It appears to be conceptually similar to the challenge of integrating relevant *in vitro* and *in silico* data.

Here the very last point on the uncertainty of the human reference dose shall be discussed more explicitly, since it addresses the highly-debated concept of thresholds for carcinogens.

3.1 Application of pragmatic deterministic standard assessment factors or data-based probabilistic assessment factors to derive a regulatory acceptable exposure level

It is well recognized that deterministic standard assessment factors applied to animal testing, i.e., no adverse effect levels (NOAELs) or benchmark dose levels (BMDLs) to derive a regulatory acceptable exposure level are very pragmatic and of unknown protection level. Thus, at first sight it may be concluded that already for quite some time we have accepted a very high level of uncertainty for defining adverse and non-adverse dose levels for humans, and consequently little scientific justification would be necessary to propose a very different approach based on defined *in silico* and *in vitro* approaches.

However, with the advancement of accessible data bases for inter-/intraspecies variation and other uncertainties, probabilistic data based assessment factors have been developed that can quantitatively inform on the uncertainties of human reference doses (e.g., Schneider et al., 2005). These are already recognized in Europe and internationally (ECHA, 2012; WHO 2014). Considering within-experimental variability, a benchmark dose with a lower and an upper confidence limit may be derived for any defined adverse effect (e.g., 10% excess risk for neoplasms). Such a benchmark dose can be represented by a probability distribution. Considering that experimental animal to human LOAEL ratios are variable over a larger set of chemicals, assessment factors accounting for animal to human differences have been developed as probability distributions. Similarly, probability distributions accounting for human to human variability and for exposure time extrapolation variability have been developed and further variability and uncertainties can be characterized in the same way. Finally, the BMD distribution can be divided by these probability distributions, i.e., probabilistic assessment factors, resulting in a probability distribution for a human reference dose providing the desired protection level, e.g., not more than 10% excess risk for neoplasms (the effect for which the BMD was modeled in the animal experiment) for, e.g., not more than 1% or 0.1% or 10^{-6}% of the human population. The 5^{th} percentile of this distribution would represent a dose that meets the desired protection level with a probability of 95%. Further non-quantifiable uncertainties should be listed, e.g., in a tabular form.

This short explanation should serve to highlight that though theoretically we may assume a threshold for any type of effect, in reality, with the use of probabilistic methods that address variability both at the experimental and extrapolation levels, we acknowledge that there will be residual hazard present at any human exposure relevant dose level. This can then be characterized, allowing the risk manager to determine the acceptable levels.

Additionally, uncertainty in terms of variance of the probabilistic human reference dose increases as we reduce the desired residual percentage of population under risk: Using the WHO APROBA tool (approximate probability assessment) it is apparent that where the P5% to P95% range for the probabilistic human reference dose is two orders of magnitude at the 1% population under risk level, this range increases to more than three orders of magnitude for a 10^{-6}% population under risk level. In other words, the quantitative precision of the human reference dose is very limited, also if calculated according to the statistically most defensible rules.

On top of this, there is qualitative uncertainty in terms of the data underlying the derivation of the probabilistic assessment factors and all issues indicated in the chapter above and, finally, ignorance, i.e., uncertainty that we are not yet aware of (Paparella et al., 2013).

Thus, overall, although we have (but still seldom use) probabilistic approaches, uncertainties around human adverse and non-adverse effect levels are still very considerable. This means that it is not an exact value that we need to predict with the new approaches, but rather that we need to describe a scientifical-

ly robust range of what is "normal" and "non-adverse", upon which we can build predictions for deviations from the "non-adverse" that are considered truly adverse. It is unlikely that these adversity margins will be precise, but they will provide a more realistic contribution to the development and assessment of a new testing strategy. It is with this perspective that new full replacement *in silico* and *in vitro* approaches appear to be more likely achievable in future (Fig. 3).

4 How could we define adversity based on defined *in silico* and *in vitro* approaches in a longer term future?

It is recognized that the real world is complex. Thinking, e.g., about aquatic environmental assessments, we are aware that many different mesocosmos may be studied in terms of temperature, pH, water hardness, organic and inorganic fractions,

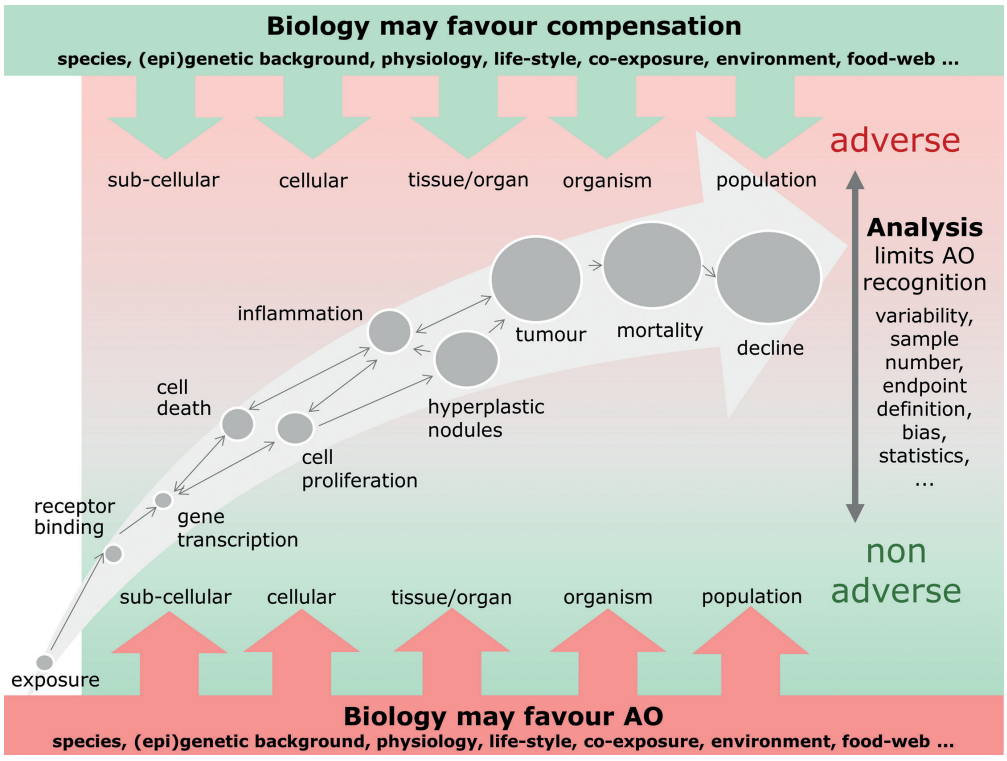

Fig. 3: Conceptual overview on uncertainties for adverse outcome recognition

The development of an adverse outcome (AO, broad arrow) evolves via key event relationships (KER, grey arrows) from events at the sub-cellular level, e.g., receptor binding and specific gene transcription, to events at the cellular level, e.g., cell death and cell proliferation, to events at the tissue or organ level, e.g., inflammation, hyperplastic nodules, tumors, to events at the organism level, e.g., morbidity, mortality and population level decline. The AO pathway approach is understood as a network with several feedback mechanisms. Biological factors, like species and strain specificities, genomic background, physiology, diet, lifestyle, co-exposures, environment, food web and others may favor the evolution of an AO (red arrows) or may favor compensatory mechanisms ameliorating the effects more towards a non-AO (green arrows), at all levels of organization. Considering that these factors represent real world variability, it may be recognized that any KER pushing the pathway towards the AO increases the probability of an AO for an individual organism with their specific biology and environment. Consequently, any of these events may be considered an adverse effect (indicated by the gradual change of background color from non-adverse green bottom to adverse red top). This recognition may allow a refined definition of adversity based on an AOP informed combination of endpoints at a level lower than the organism level. Note that for AOP development the AO may be defined at any level of organization (from sub-cellular to population) and the AOP shall be as generic, i.e., species and chemical independent, as reasonably practical (OECD, 2013). Finally, the experimental analysis limits the recognition of the events, inter alia due to control variability, sample number, endpoint definition, potential for bias and statistics, which introduces further variability and complexity into the definition of adverse effects. Note that the grey shaded AO arrow broadens from the sub-cellular to the population level, indicating that one could expect that as the experimental variability is likely to increase with the complexity of the system, the simpler *in vitro* (MIE) test systems would provide the added benefit of relatively reduced uncertainty compared to the more complex *in vivo* systems. Analyzing all these uncertainties and complexities of current approaches undertaken shall serve the development of new and better approaches.

sunlight, plant and animal species, population sizes, food web and others. One mesocosmos is not necessarily representative for another mesocosmos and this is one of the main reasons why regulators usually prefer to estimate a bioconcentration factor (BCF) via a reductionist laboratory fish test rather than a trophic magnification factor (TMF) via a complex mesocosmos study (see e.g., ECHA, 2014).

Similarly, human beings are complex and variable over time and variable between each other. Adverse outcomes, e.g., in terms of tumor development, depend on genetic background, physiology, pre-existing disease status, life-style, life-stage of exposure and co-exposure. There are many good examples of human variability in adverse responses to chemical exposure. Taking smoking as an example (smoke contains genotoxic and non-genotoxic components), it is known that only about 20% of smokers develop cancer (all cancers combined), of which there appear to be clear gender differences, i.e., 16% for women vs 22% for men (NCI, 2016; Parkin, 2011), and acetylation speed (fast versus slow acetylation) is a major mechanistic variable. Indeed, the association between breast cancer and smoking is polymorphism-status dependent, in female slow acetylators, the association is 4.4 times higher (Woodson et al., 1999). The association between specific polymorphisms and cancer has been noted in heavy smokers participating in two large cancer chemoprevention studies, CARET and ATBC (Doherty et al., 2013). As another example, not all workers exposed to aromatic amines develop bladder tumors (Antonova et al., 2015).

This means that the result at the level of an individual organism does not tell us everything we want to know – especially since we want to protect many individuals. In epidemiology one answer is to aim for large group sizes, but it is also well known how sensitive group selection is for the outcome of the analysis. So called "confounders" represent part of the real life human variability that epidemiologists need to minimize for their analysis and conclusion. We experience a similar situation with animal testing in toxicology in that also in animal tests tumor incidence increase may depend on species, strain, sex, age, and feed amongst other factors, and one impractical, costly and unethical answer may be animal testing of multiple species for multiple generations in multiple environmental conditions. However, though some variability can be experimentally assessed and addressed – in practice with animal testing – animal response variability is deliberately reduced by using inbred strains and standardization of housing and feed since we aim to detect small differences between groups.

In summary, we are already intentionally working with reductionist methods in ecotoxicology, in epidemiology and with animal testing in human health toxicology. This is because we are not interested in chemical concentrations leading to adverse outcomes in any specific real life ecosystems or in any specific real life human population or in any specific laboratory animal strain/condition. But we are interested in the chemical concentrations that may lead to adverse outcomes in a small number of ecosystems, humans or laboratory animal strains under any potentially realistic conditions. This is true especially with increasing interest in protecting the environment from subtle, long term effects and with aiming to reduce the exposure of sensitive human (sub)populations to minimal/low or no effect levels. At this point, rather than increasing the complexity and number of species, strains and individuals under analysis (highly impractical), conceptually we may be better off examining the upstream levels in the AOP, i.e., the cellular and sub-cellular levels, to find an optimum combination of relevant endpoints that are less complex, i.e., more reductionist and more reproducible, but that are mechanistically convincing, if affected, and can be shown to substantially contribute to the increased likelihood of adverse outcomes on the level of any individual organism or population (Fig. 3). Chepelev and colleagues (2015) propose using the key events that are more sensitive for the MoA, at the tissue level, and the closest to the adverse outcome. This may even allow a protection level that is higher than any standard animal test, i.e. beyond the recognized issue of potential MoA differences, always limited *inter alia* by the number of animals and background variability.

Scientific and regulatory communities appear to be increasingly moving towards this perspective. In a consensus statement on the identification of endocrine disrupting chemicals, the WHO definition of endocrine disruptors was recently re-interpreted where an effect in the "intact organism" is now understood to mean that the effect would occur *in vivo*, either observable in a test animal system, epidemiologically or clinically. And, most importantly, that it "does not necessarily mean that the adverse effect has to be demonstrated in an intact test animal, *but may be shown in adequately validated alternative test systems predictive of adverse effects in humans and/or wildlife*" (Solecki et al., 2016, emphasis added).

With the knowledge of a potential downstream *in vivo* adverse carcinogenic related consequence to a specified mechanism and MoA we may define adversity on a MoA effect level that is subtler than that observed in "black box" *in vivo* studies, such that we actually cannot observe it in any randomly selected variant of a standard animal test. But is that wrong in principle? Or does it just represent a break with current practice?

Current toxicological practice is already in the throes of change, with the rapid expansion of work on AOPs and IATAs under the auspices of the OECD (e.g., Jacobs et al., 2016 and references therein, http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm), and the consequent advancement in our mechanistic understanding of AOPs.

As outlined in Sections 2 and 3 above, the solution to this discussion towards a refined understanding of adversity needs a systematic analysis of the reliability and relevance, complexity and ambiguity of actual *in vivo* reference data and assessments including a review of the amount of pragmatism necessary for defining adversity based on animal organism level results. The results of such an analysis can then be compared to the uncertainties and necessary pragmatism to define adversity based on a molecular and cellular level response. A potential frame for a systematic analysis with regard to carcinogenicity is explored in Section 3 and Table 2.

5 How could we regulate with a new *in silico* and *in vitro* based adversity concept in a longer term future?

We already have several *in silico* and *in vitro* methods that can test for the molecular, sub-cellular or cellular events upstream of adverse apical effects, but only the latter, i.e., effects at organism level, are currently accepted for classification, e.g., for carcinogenicity, mutagenicity, reproductive toxicity, and specific target organ toxicity. However, similar MoAs may lead to various apical adverse effects: For example, receptor mediated effects and cellular stress response pathways may trigger organ toxicity, carcinogenicity (see e.g., figures for multistep carcinogenesis in Jacobs et al., 2016), but also developmental toxicity and other toxicities. Therefore, we may consider redefining adversity on a molecular and cellular level (Fig. 3), then translating the *in vitro* BMD to a corresponding *in vivo* dose (i.e., quantitative kinetic *in vitro* to *in vivo* extrapolation modeling (QIVIVE); McNally and Loizou, 2015; Yoon et al., 2015), and finally classifying according to potency. In this way, we might distinguish substances with medium versus very low dose adverse effect levels, categorize and regulate them according to the *in vitro* potency categories, regardless of what the actual standard "adverse apical effect" is (It is in any case an effect that we want to protect against). We may then add "for information" a probability for the actual carcinogenicity/ mutagenic/reproductive toxicity/specific target organ toxicity (CMR/ STOT), etc. classes, but we would regulate on the basis of the *in vitro* MoA hazard class and categories.

Ultimately, considering the concept and international work along the NIH 21st century toxicology initiative, it may be useful to consider introducing a Globally Harmonized System (GHS) "*in vitro* MoA hazard class", which could develop over time to include more and more critical MoAs and key events, as an addition to current GHS classes. Such approaches are currently being explored by the UN Sub-Committee of Experts on the Globally Harmonized System of Classification and Labelling of Chemicals in the field of skin corrosion/ irritation[6] and this would be a useful foundation, potentially for further applications. In order to serve for regulatory use, such a new hazard class would need to be scientifically robust enough to facilitate definitive regulatory decision-making on necessary risk mitigation and risk management measures, including restrictions on the market, as is the case for example for biocides and pesticides in Europe.

Furthermore, one may be able to use *in vitro* data as hazard alerts upon which we may also be able to derive a 'probably acceptable human effect level based on *in vitro* approaches', in the sense that the new approaches may "predict likely safe exposures for specific toxicity pathways, rather than organ toxicity per se" (Dowes and Foster, 2015) (Fig. 4).

The need to start considering such developments may be substantiated by the following lines of thought:

- From the environmental and human health perspective it appears desirable to increase the protection level for chemicals to cover subtle long term environmental effects as well as a high percentile of the human (sub)population. This may not be easily achievable with standard animal tests, but may lead to a refined understanding of adversity along the discussion in Sections 3 and 4 of this article..
- Testing throughput is increasing with the continuing transition towards defined *in silico* and *in vitro* approaches. With more and more chemicals being tested, also more chemicals may be expected to appear to be "positive" for the one or other critical effect. In this case, further sub-categorizing between "positives" may be necessary in order to focus regulation on the most critical chemicals (Jacobs et al., 2016).
- Testing an increasing amount of substances with the same *in vitro* approach may increase data availability from similar testing strategies and thus allow a better comparison of chemicals and their potency (compared to the current situation where we have high variability of – largely animal – test designs).
- With increasing AOP network development and *in vitro* test availability, a refined selection of test methods may be needed to increase efficacy of testing. Targeting *in vitro* testing to key events and key event relationships with the most critical and far reaching influence on the AOP network may prove important for this aim.
- Testing for the most conserved key events and key event relationships shall also support cross-species extrapolation (Groh et al., 2015), which has the potential to reduce additional regulatory testing needs where equivalent concordance is scientifically plausible.

In considering such an approach, it may be helpful to recall that currently we classify only for the critical positive effects, and then only in situations where sufficient reliable data are available. We are also aware that there are effects like respira-

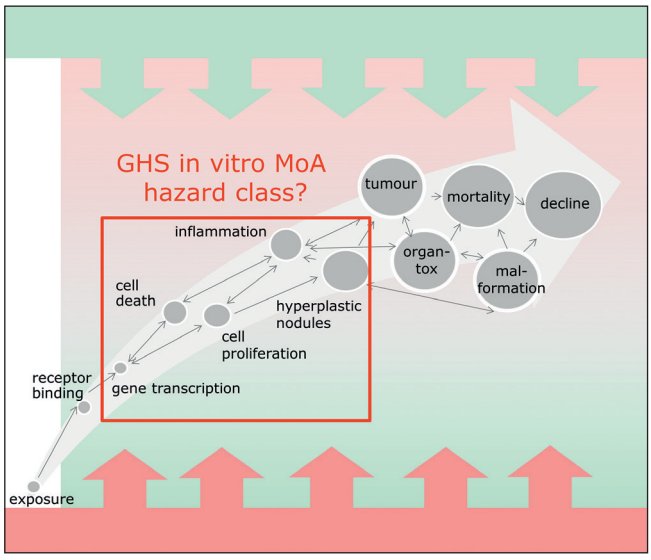

Fig. 4: Conceptual consideration for a new GHS *in vitro* toxicity class

[6] UN 2016: http://bit.ly/2nVilOZ; http://bit.ly/2mLKY11

tory sensitization, for which no standard animal test exists, and that for most chemicals an ample, reliable and relevant data set is neither available nor practically achievable. Furthermore, we also do not expect that animal testing results inform us on the exact type of adverse outcome that will occur in humans: Whatever specific malformation or embryotoxicity is observed in standard animal tests, it is interpreted as potential for any human malformation or embryotoxicity, and the same is true for carcinogenic effects and sites, onset and progression of neoplasms. Indeed, there are no suitable animal models for many human cancers. The exact human adverse outcome is neither predicted nor of interest. It is sufficient to know that we have a problem, to understand the concentration/dose level at which it occurs, and whether this is a realistic exposure scenario or not. Similarly, where environmental toxicity is estimated, the exact adverse environmental outcome in terms of affected species and ecosystems is not and cannot be predicted due to the complexity and variability of the ecosystems. This is also mirrored in the GHS classification approach that provides a single class for all (aquatic) environmental effects and differentiates for potency.

Taking a longer-term perspective, it is likely that further options may arise and be discussed to facilitate the evolution of chemical regulations so that they can reap a greater benefit from the new approaches and better address the purpose and needs of environmental and human health protection goals. Indeed, the development and application of a few, well-understood key event relationships within robust testing strategies for regulatory applications to many chemicals may better and more comprehensively serve hazard assessment and ultimately risk management than the testing of relatively few chemicals in low throughput and "black box" animal *in vivo* tests.

6 Conclusion

A systematic analysis of the uncertainties and complexity of the seemingly "golden" animal based testing and assessment standards is needed. This will allow agreement upon what appropriate reference results should be, and agreement upon the weight that "gold" standard animal testing reference data should have when assessing the relevance and utility of new approaches compared to appropriately selected mechanistic and human data.

Here, the use of the OECD template for defined *in vitro* and *in silico* approaches is utilized and explored as a systematic frame for analysis of standard animal based approaches, exemplarily for the standard *in vivo* carcinogenicity testing and assessment. The format appears suitable to support, structure and substantiate the discussion on the complexity and potential multitude of approaches for integrating and interpreting data from standard animal testing approaches, which ultimately appears to be conceptually similar to the challenge of integrating relevant *in vitro* and *in silico* data. On this basis, complexity and uncertainties of current standard *in vivo* approaches appear considerable and continue the call for a paradigm change in testing and assessment of carcinogenicity.

A concept for concluding on the sufficient performance of new defined *in silico* and *in vitro* approaches compared to the current standard *in vivo* approaches is suggested, giving more weight to carefully selected mechanistic and human reference data and information. Moreover, in order to succeed with the development and application of new defined approaches, also our understanding of what constitutes an adverse effect needs evolution: On the one hand, there is a high variability of real life chemical exposures and responses to exposure, and it is not feasible to test and assess all potentially relevant situations. On the other hand, any experiment and analysis is limited in its ability to recognize adversity due to animal or sample number for sufficient power to detect a real effect, as well as the appropriate use of statistics, and so on.

As part of the resolution of these problems, there is growing momentum now to change this approach, such that adversity may be defined within the cellular and sub-cellular context as an increase in the probability for an adverse outcome at organism and population level. It may be that such adversity is not even recognized within a random protocol variant of an animal test, but that does not mean that it could not be considered more reliable and relevant in the future.

Finally, chemical regulations may need to evolve to be able to fully embrace the opportunities that new 21st century toxicology approaches are starting to offer. This may include possibilities to evolve GHS in a number of ways, including, for example, by introducing a class of differentiated categories for critical *in vitro* assessed pathways. This may support the testing and regulation of a much higher number of chemicals via the use of *in silico* and *in vitro* testing results on the basis of well understood key pathways. In this article some perspectives are put forward for these aspects in order to stimulate discussion and to progress such regulatory toxicology evolution.

References

Adami, H. O., Berry, S. C., 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. https://doi.org/10.1093/toxsci/kfr113

Alden, C. L., Lynn, A., Bourdeau, A. et al. (2011). A critical review of the effectiveness of rodent pharmaceutical carcinogenesis testing in predicting for human risk. *Vet Pathol 48*, 772-784. https://doi.org/10.1177/0300985811400445

Andersson, N. (2015). Regulatory perspective on non-monotonic dose-response curves and low dose effects. *Toxicol Lett 238, Suppl,* S53. https://doi.org/10.1016/j.toxlet.2015.08.148

Annys, E., Billington, R., Clayton, R. et al. (2014). Advancing the 3Rs in regulatory toxicology – carcinogenicity testing: Scope for harmonisation and advancing the 3Rs in regulated sectors of the European Union. *Regul Toxicol Pharmacol 69*, 234-242. https://doi.org/10.1016/j.yrtph.2014.04.009

Antonova, O., Toncheva, D. and Grigorov, E. (2015). Bladder cancer risk from the perspective of genetic polymorphisms in the carcinogen metabolizing enzymes. *J BUON 20*, 1397-1406.

Baldeshwiler, A. M. (2003). History of FDA good laboratory practices. *Qual Assur J 77*, 157-161. https://doi.org/10.1002/qaj.228

Bal-Price, A., Crofton, K. M., Leist, M. et al. (2015). International STakeholder NETwork (ISTNET): Creating a developmental neurotoxicity (DNT) testing road map for regulatory purposes. *Arch Toxicol 89*, 269-287. https://doi.org/10.1007/s00204-015-1464-2

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. https://doi.org/10.14573/altex.2012.1.003

Berenblum, I. (1941). The mechanism of carcinogenesis. A study of the significance of cocarcinogenic action and related phenomena. *Cancer Res 1*, 807f.

Billington, R., Lewis, R. W., Mehta, J. M. and Dewhurst, I. (2010). The mouse carcinogenicity study is no longer a scientifically justifiable core data requirement for the safety assessment of pesticides. *Crit Rev Toxicol 40*, 35-49. https://doi.org/10.3109/10408440903367741

Blettner, M., Heuer, C. and Razum, O. (2001). Critical reading of epidemiological papers. A guide. *Eur J Public Health 11*, 97-101. https://doi.org/10.1093/eurpub/11.1.97

Boobis, A. R., Cohen S. M., Dellarco V. et al. (2006). IPCS framework for analyzing the relevance of a cancer mode of action for humans. *Crit Rev Toxicol 36*, 781-92. https://doi.org/10.1080/10408440600977677

Borak, J. and Sirianni G. (2005). Hormesis: Implications for cancer risk assessment. *Dose Response 3*, 443-451. https://doi.org/10.2203/dose-response.003.03.011

Chepelev, N. L., Moffat, I. D., Labib, S. et al. (2015). Integrating toxicogenomics into human health risk assessment: Lessons learned from the benzo[a]pyrene case study. *Crit Rev Toxicol 45*, 44-52. https://doi.org/10.3109/10408444.2014.973935

COC (2016). Alternatives to the 2-year Bioassay: Committee on Carcinogenicity of Chemicals in Food, Consumer Products and the Environment, Statement COC/G07, 2 February 2016. https://www.gov.uk/government/publications/alternatives-to-the-2-year-bioassay

Cook, J., Hewett, C. and Hieger, I. (1933). The isolation of a cancer-producing hydrocarbon from coal tar. Parts I, II, and III. *J Chem Soc 24*, 395-405. https://doi.org/10.1039/jr9330000395

CRD – Chemicals Regulation Directorate, Health & Safety Executive, UK (2013). Investigation of the state of the art on identification of appropriate reference points for the derivation of health-based guidance values (ADI, AOEL and AAOEL) for pesticides and on the derivation of uncertainty factors to be used in human risk assessment. *EFSA Supporting Publications 10*, EN-413. http://onlinelibrary.wiley.com/doi/10.2903/sp.efsa.2013.EN-413/abstract

Creasy, D., Bube, A., de Rijk, E. et al. (2012). Proliferative and nonproliferative lesions of the rat and mouse male reproductive system. *Toxicol Pathol 40, Suppl 6*, 40S-121S. https://doi.org/10.1177/0192623312454337

Demetrius, L. (2006). Aging in mouse and human systems: a comparative study. *Ann N Y Acad Sci 1067*, 66-82. https://doi.org/10.1196/annals.1354.010

Doherty, J. A., Sakoda, L. C., Loomis, M. M. et al. (2013). DNA repair genotype and lung cancer risk in the beta-carotene and retinol efficacy trial. *Int J Mol Epidemiol Genet 4*, 11-34.

Downes, N. and Foster, J. (2015). Regulatory forum opinion piece: Carcinogen risk assessment: The move from screens to science. *Toxicol Pathol 43*, 1064-1073. https://doi.org/10.1177/0192623315598578

Dutta, S. and Sengupta, P. (2016). Men and mice: Relating their ages. *Life Sci 152*, 244-248. https://doi.org/10.1016/j.lfs.2015.10.025

EC (2013). Thresholds for Endocrine Disrupters and Related Uncertainties. Report of the Endocrine Disrupters Expert Advisory Group. European Commission.

ECHA (2012). Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.19 Uncertainty analysis.

ECHA (2014). Guidance on Information Requirements and Chemical Safety Assessment Part C: PBT/vPvB assessment.

ECHA (2016). Topical Scientific Workshop – New Approach Methodologies in Regulatory Science. http://echa.europa.eu/news-and-events/events/event-details/-/journal_content/56_INSTANCE_DR2i/title/topical-scientific-workshop-new-approach-methodologies-in-regulatory-science

EEA (2010). Prudent Precaution? Experiences with the Precautionary Principle, 2000-2010. http://www.umweltbundesamt.at

Ennever, F. K. and Lave, L. B. (2003). Implications of the lack of accuracy of the lifetime rodent bioassay for predicting human carcinogenicity. *Regul Toxicol Pharmacol 38*, 52-57. https://doi.org/10.1016/S0273-2300(03)00068-0

Gaylor, D. W. (2005). Are tumor incidence rates from chronic bioassays telling us what we need to know about carcinogens? *Regul Toxicol Pharmacol 41*, 128-133. https://doi.org/10.1016/j.yrtph.2004.11.001

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

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

Groh, K. J., Carvalho, R. N., Chipman, J. K. et al. (2015). Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology. *Chemosphere 120*, 764-777. https://doi.org/10.1016/j.chemosphere.2014.09.068

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

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

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

IRGC (2005). *IRGC White Paper No1 "Risk Governance – Towards an Integrative Approach"*. Geneva: IRGC. ISBN 978-2-9700772-2-0. http://www.irgc.org

Jacobs, A. C. and Hatfield, K. P. (2013). History of chronic toxicity and animal carcinogenicity studies for pharmaceuticals. *Vet Pathol 50*, 324-333. https://doi.org/10.1177/0300985812450727

Jacobs, M. N., Colacci, A., Louekari, K. et al. (2016). International regulatory needs for development of an IATA for non-genotoxic carcinogenic chemical substances. *ALTEX 33*, 359-392. https://doi.org/10.14573/altex.1601201

Jaworska, J., Harol, A., Kern, P. S. and Gerberick, G. F. (2011). Integrating non-animal test information into an adaptive testing strategy – skin sensitization proof of concept case. *ALTEX 28*, 211-225. https://doi.org/10.14573/altex.2011.3.211

Leist, M., Hasiwa, N., Rovida, C. et al. (2014). Consensus report on the future of animal-free systemic toxicity testing. *ALTEX 31*, 341-356. https://doi.org/10.14573/altex.1406091

Lewis, D. F. V. (2002). Pesticides and P450 induction: Gender implications. In M. N. Jacobs and B. Dinham (eds.), *Silent Invaders, Pesticides Livliehoods and Women's Health*. London: Publ Zed books.

Linkov, I., Bates, M. E., Trump, B. D. et al. (2013). For nanotechnology decisions, use decision analysis. *Nano Today 8*, 5-10. https://doi.org/10.1016/j.nantod.2012.10.002

Maertens, A., Anastas, N., Spencer, P. J. et al. (2014). Green toxicology. *ALTEX 31*, 243-249. https://doi.org/10.14573/altex.1406181

Mahadevan, B., Snyder, R. D., Waters, M. D. et al. (2011). Genetic toxicology in the 21st century: Reflections and future directions. *Environ Mol Mutagen 52*, 339-354. https://doi.org/10.1002/em.20653

McNally, K. and Loizou, G. D. (2015). A probabilistic model of human variability in physiology for future application to dose reconstruction and QIVIVE. *Front Pharmacol 6*, 213. https://doi.org/10.3389/fphar.2015.00213

McNally, K., Cotton, R., Hogg, A. and Loizou, G. (2015). Reprint of PopGen: A virtual human population generator. *Toxicology 332*, 77-93. https://doi.org/10.1016/j.tox.2015.04.014

Meek, M. E., Boobis, A., Cote, I. et al. (2014). New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. *J Appl Toxicol 34*, 1-18. https://doi.org/10.1002/jat.2949

NAS (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington: The National Academies Press.

NAS (2015). Reproducibility Issues in Research with Animals and Animal Models: Workshop in Brief. National Academies of Sciences. http://www.nap.edu/read/21835/

NCI (2016). Cigarette Smoking: Health Risks and How to Quit: National Cancer Institute. http://www.cancer.gov/about-cancer/causes-prevention/risk/tobacco/quit-smoking-pdq#section/_10

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 34*. ENV/JM/MONO(2005)14

OECD (2007). Guidance document on the validation of the quantitative structure activity relationship s (QSAR) models. *OECD Series on Testing and Assessment 69*. ENV/JM/MONO(2007)2

OECD (2012a). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins, part 2. *OECD Series on Testing and Assessment 168*. ENV/JM/MONO(2012)10/PART2

OECD (2012b). Guidance doucment 116 on the conduct and design of chronic toxicity and carcinogenicity studies, supporting test guidelines 451, 452 and 453. *OECD Series on Testing and Assessment 116*. ENV/JM/MONO(2011)47

OECD (2013). Guidance document on developing and assessing adverse outcome pathways. *OECD Series on Testing and Assessment 184*. ENV/JM/MONO(2013)6

OECD (2014). Guidance on grouping of chemicals, second edition. *OECD Series of Testing and Assessment 194*. ENV/JM/MONO(2014)4

OECD (2015) Report of the Workshop on a Framework for the Development and Use Of Integrated Approaches to Testing and Assessment. *OECD Series of Testing and Assessment 215*. ENV/JM/MONO(2015)22

OECD (2016). Guidance document on the reporting of defined approaches to be used within integrated approaches to testing and assessment. *Series of Testing and Assessment 255*. ENV/JM/MONO (2016)28

Paparella, M., Daneshian, M., Hornek-Gausterer, R. et al. (2013). Uncertainty of testing methods – what do we (want to) know? *ALTEX 30*, 131-144. https://doi.org/10.14573/altex.2013.2.131

Parkin, D. M. (2011). 1. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br J Cancer 105, Suppl 2*, S2-5. https://doi.org/10.1038/bjc.2011.474

Proctor, D. M., Gatto, N. M., Hong S. J. and Allamneni, K. P. (2007). Mode-of-action framework for evaluating the relevance of rodent forestomach tumors in cancer risk assessment. *Toxicol Sci 98*, 313-326. https://doi.org/10.1093/toxsci/kfm075

Renn, O., Klinke, A. and van Asselt, M. (2011). Coping with complexity, uncertainty and ambiguity in risk governance: A synthesis. *Ambio 40*, 231-246. https://doi.org/10.1007/s13280-010-0134-0

Rudén, C. (2001). The use and evaluation of primary data in 29 trichloroethylene carcinogen risk assessments. *Regul Toxicol Pharmacol 34*, 3-16. https://doi.org/10.1006/rtph.2001.1482

Sailaukhanuly, Y., Zhakupbekova, A., Amutova, F. and Carlsen, L. (2013). On the ranking of chemicals based on their PBT characteristics: Comparison of different ranking methodologies using selected POPs as an illustrative example.

Chemosphere 90, 112-117. https://doi.org/10.1016/j.chemosphere.2012.08.015

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

Schneider, K., Hassauer, M., Ottmanns, J. et al. (2005). Uncertainty analysis in workplace effect assessment. Federal Institute for Occupational Safety and Health. Research Project F 1824, F1825, F 1826. http://www.baua.de/nn_21712/en/Publications/Expert-Papers/Gd36.html

Schneider, K., Schwarz, M., Burkholder, I. et al. (2009). "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol Lett 189*, 138-144. https://doi.org/10.1016/j.toxlet.2009.05.013

Schroeder, A. L., Ankley, G. T., Houck, K. A. and Villeneuve, D. L. (2016). Environmental surveillance and monitoring – the next frontiers for high-throughput toxicology. *Environ Toxicol Chem 35*, 513-525. https://doi.org/10.1002/etc.3309

Sengupta, P. (2013). The Laboratory Rat: Relating Its Age With Human's. *Int J Prev Med 4*, 624-630.

Solecki, R., Kortenkamp, A., Bergman, Å. et al. (2017) Scientific principles for the identification of endocrine-disrupting chemicals: a consensus statement. *Arch Toxicol 91*, 1001-1006. https://doi.org/10.1007/s00204-016-1866-9

Testai, E. (2015). Dose-response relationship: Monotone vs non-monotone curve. *Toxicol Lett 238, Suppl 2*, S51. https://doi.org/10.1016/j.toxlet.2015.08.144

U. S. EPA (2013). State of the Science Evaluation: Nonmonotonic Dose Responses as They Apply to Estrogen, Androgen, and Thyroid Pathways and EPA Testing and Assessment Procedures. In *State of the Science: Nonmonotonic Dose Responses V7*.

Van Oosterhout, J. P., Van der Laan, J. W., De Waal, E. J. et al. (1997). The utility of two rodent species in carcinogenic risk assessment of pharmaceuticals in Europe. *Regul Toxicol Pharmacol 25*, 6-17. https://doi.org/10.1006/rtph.1996.1077

Vandenberg, L. N., Welshons, W. V., Vom Saal, F. S. et al. (2014). Should oral gavage be abandoned in toxicity testing of endocrine disruptors? *Environ Health 13*, 46. https://doi.org/10.1186/1476-069X-13-46

WHO (2014). Guidance document on evaluating and expressing uncertainty in hazard characterization. *IPCS Harmonization Project Document 11*.

WHO (2015). Guidance document for WHO monographers and reviewers. WHO/HSE/FOS/2015.1

Woodson, K., Ratnasinghe, D., Bhat, N. K. et al. (1999). Prevalence of disease-related DNA polymorphisms among participants in a large cancer prevention trial. *Eur J Cancer Prev 8*, 441-447. https://doi.org/10.1097/00008469-199910000-00010

Yamagiwa, K. and Ichikawa, K. (1918). Experimental study of the pathogenesis of carcinoma. *J Cancer Res 3*, 1-29.

Yoon, M., Blaauboer, B. J. and Clewell, H. J. (2015). Quantitative in vitro to in vivo extrapolation (QIVIVE): An essential element for in vitro-based risk assessment. *Toxicology 332*, 1-3. https://doi.org/10.1016/j.tox.2015.02.002

Young, R. M., Hardy, I. R., Clarke, R. L., et al. (2009). Mouse models of non-Hodgkin lymphoma reveal Syk as an important therapeutic target. *Blood 113*, 2508-2516. https://doi.org/10.1182/blood-2008-05-158618

Zeise, L., Bois, F. Y., Chiu, W. A. et al. (2013). Addressing human variability in next-generation human health risk assessments of environmental chemicals. *Environ Health Perspect 121*, 23-31. https://doi.org/10.1289/ehp.1205687

Conflict of interest

The authors declare to not have any conflict of interest with regard to the content of this article.

Acknowledgment

The authors would like to thank Dr Robin Foster of the UK Health and Safety Executive for helpful comments regarding the GHS and aspects of its current work.

Correspondence to

Martin Paparella, MS(Tox), PhD
Chemicals and Biocides Unit
Environment Agency Austria
Spittelauer Lände 5
1090 Vienna, Austria
Phone: +43 1 31304 3407
e-mail: martin.paparella@umweltbundesamt.at